Systems Biology of Plant Molecular Networks:

From Networks to Models

Felipe Leal Valentim

Thesis committee

Promotor
Prof. Dr G.C. Angenent
Personal chair at the Laboratory of Molecular Biology
Wageningen University

Co-promotor
Dr A.D.J. van Dijk
Senior scientist, Bioscience
Plant Research International
Wageningen

Other members
Prof. Dr D. de Ridder, Wageningen University
Dr J.D. Stigter, Wageningen University
Prof. Dr G. Coupland, Max Planck Institute for Plant Breeding Research, Köln, Germany
Prof. Dr B. Snel, Utrecht University

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences (EPS)

Systems Biology of Plant Molecular Networks:

From Networks to Models

Felipe Leal Valentim

Thesis
submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. Dr M.J. Kropff,
in the presence of the
Thesis committee appointed by the Academic Board
to be defended in public
on Monday 19th of January 2015
at 4 p.m. in the Aula.

Felipe Leal Valentim
Systems Biology of Plant Molecular Networks:From Networks to Models
136 pages

PhD thesis, Wageningen University, Wageningen, NL (2015)
With references, with summaries in Dutch and English

ISBN 978-94-6257-217-1

Contents

Chapter 1

Systems biology of molecular networks controlling Arabidopsis plant reproduction: from networks to predictive models 7

Chapter 2

A quantitative and dynamic model of the Arabidopsis flowering time gene regulatory Network 23

Chapter 3

Predicting the effect of SNPs on flowering time 57

Chapter 4

Interactome-wide prediction of protein-protein binding sites reveals effects of protein sequence variation in Arabidopsis thaliana 85

Chapter 5

Concluding remarks 117

Summary 127

Samenvatting 130

Cover design: Proefschriftmaken.nl || Uitgeverij BOXPress

Printed by: Proefschriftmaken.nl || Uitgeverij BOXPress

Published by: Uitgeverij BOXPress, ’s-Hertogenbosch

Chapter 1

Systems biology of molecular networks controlling Arabidopsis plant reproduction: from networks to predictive models

Felipe Leal Valentim

Gene regulatory networks controlling plant reproduction

Developmental processes are controlled by tightly coordinated networks of regulators, known as gene regulatory networks (GRNs) that activate and repress gene expression within a spatial and temporal context. In *Arabidopsis thaliana*, the key components of GRNs, e.g. those controlling major processes in plant reproduction such as the floral transition and floral organ identity specification, were first identified in loss of function mutants that affect these processes [1]. The interactions between these regulators later began to be revealed through genetic analyses, resulting in the first, mostly linear, GRN snapshots. These were augmented and detailed by reverse genetics, analysis of protein-protein interactions and expression studies in wild type and mutant plants, resulting in a hierarchical GRN in which master regulators target a subset of genes directing downstream processes [1] (Fig. 1). In this chapter, I introduce different type of systems biology approaches that can be used to study those networks. Subsequently, the scope and outline of the remaining of this thesis is presented, in which we make use of these systems biology approaches.

From information to understanding, from networks to predictive models

We have been witnessing a revolution in plant biology pivoted by scientific advancements in the 'omics' technologies. We are now starting to systematically cataloguing the molecules and their interactions within a cell, tissue, organ or organism; this for several growth conditions, developmental stages or genetic backgrounds. On top of the genomes, a vast amount of data are being generated through phenomics, transcriptomics, proteomics, interactomics and protein-DNA binding profiling (e.g. by ChIP-seq), as well as through emerging high throughput technologies for the elucidation of the epigenome .Yet, because of and despite the wealth of data, there is a clear need to understand how these molecules regulate complex traits. In this direction, a first goal of systems biology is to provide systems-wide representation, or systems-wide snapshots, in which the relationships and interactions between the molecules, as well as their relevant features, are comprehensively represented. By representing the snapshots over time, much can be said about the dynamics of the system.

A system can be defined at different levels of biological organisation, from molecules to ecosystems. For instance, these levels can be represented by (i) molecular signalling pathways at a subcellular level, (ii) networks of physiological process at the cellular level, (iii) plant growth and development at an individual level, or (iv) genetic variation among individuals within species at the population level [2]. Once the molecules of a biological system have been comprehensively represented, as well as their interactions, relationships and relevant features, it is then interesting to interrogate this representation in order to understand the system at a mechanistic level. For simple, linear systems, this would be a straightforward and intuitive task, but for complex non-linear biological systems, analytical tools are needed. For that, systems biology offers two approaches; 1) modelling concepts

and techniques that can be applied to integrate the different levels of organization into predictive models; or more directly but not simpler, 2) bioinformatics techniques that can be used to perform multidimensional data analysis [3]. The challenges addressed in this thesis concern the use of both dynamic modelling and bioinformatics approaches for multidimensional data analysis. We focus on applying these systems biology approaches for studying and modelling gene regulatory networks underlying plant reproduction processes (see Fig. 1).

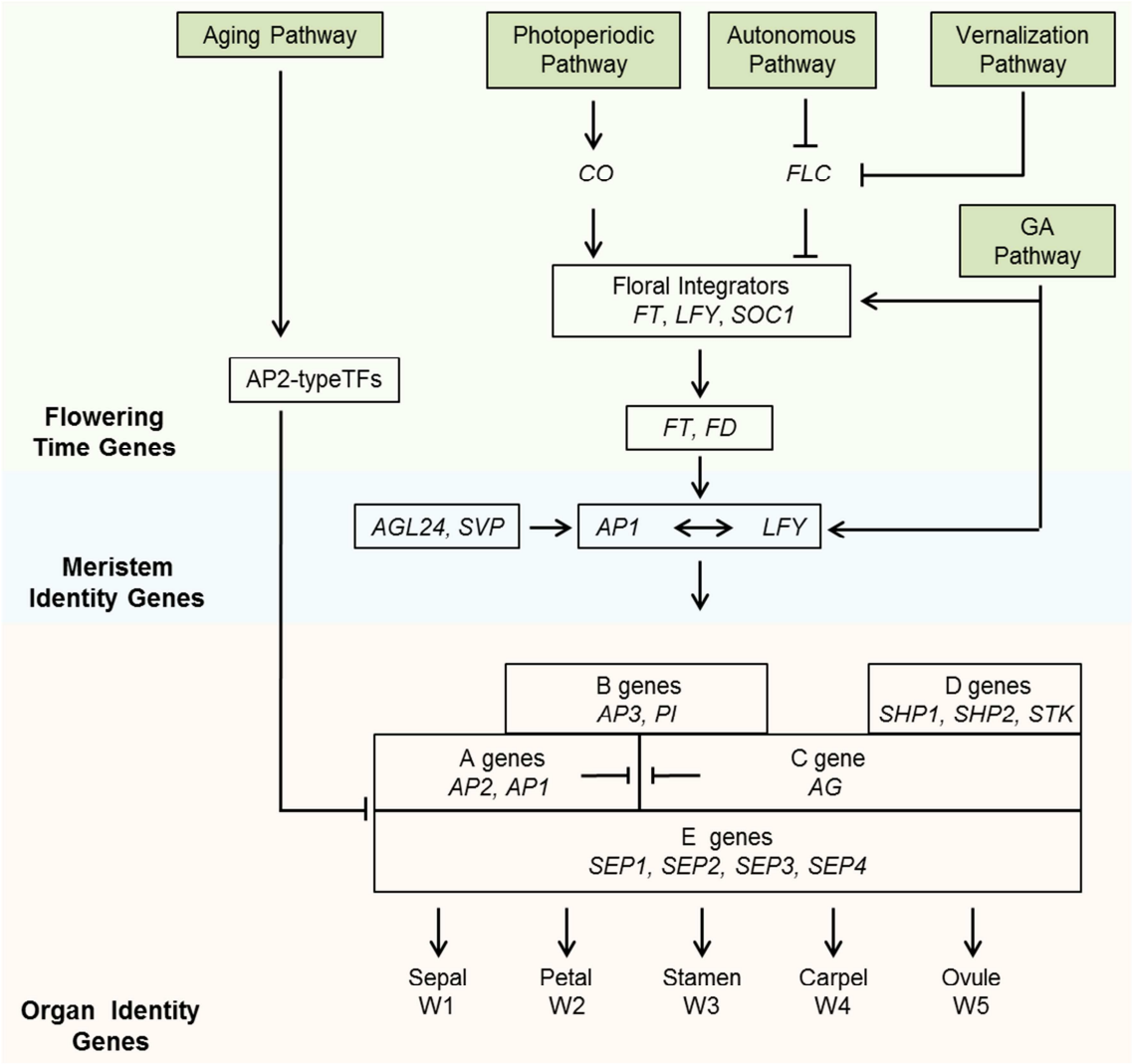


Figure 1: Linear GRN of genes controlling flower formation. The switch from vegetative to reproductive phase is triggered by endogenous and environmental stimuli; some key regulators are illustrated here. The vernalisation and photoperiod signals converge at the flowering regulator genes *FLC* and *CO*, respectively, that antagonistically regulate the floral integrator genes, *FT* in the leaf and *LFY* and *SOC1* in the at the shoot apical meristem (SAM). The floral integrators activate the meristem identity genes *AP1* and *LFY*, subsequently leading to activation of the ABCDE class genes, specifying organ identity [4]. The endogenous aging pathway involves micro RNAs (miRNAs). Arrows indicate activation, blocked lines indicate repression, left-right arrows indicate a positive feedback loop. Abbreviations: AGL24, AGAMOUS LIKE 24; AP1, APETALA1; CO, CONSTANS; FD, FLOWERING LOCUS D; FT, FLOWERING LOCUS T; LFY, LEAFY; STK, SEEDSTICK; SEP, SEPALLATA; SHP, SHATTERPROOF; SOC1, SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1;

SVP, SHORT VEGETATIVE PHASE; SPL, SQUAMOSA PROMOTER BINDING PROTEIN-LIKE; W1; whorl 1, W2; whorl 2, W3; whorl 3, W4; whorl4.

Modelling gene regulatory networks

An important aspect of understanding GRNs is how perturbations in one part of the network are transmitted to other parts of the network, and ultimately how this results in changes in phenotype. The latest versions of Arabidopsis GRNs as presented by Fig. 1 involve highly-connected, non-linear networks [5]. Complex gene regulatory mechanisms underlie these processes, including transcription factors, microRNAs, movable factors, hormones and chromatin modifying proteins. Given this complexity, it is not possible to predict the effect of gene perturbations on e.g. flowering time in an intuitive way. Therefore, mathematical modelling plays an important role in providing a quantitative understanding of GRNs.

One of the pioneering models for cell-fate determination during the formation of floral organ primordia in Arabidopsis provided insights into the semi-quantitative relationships between the genes in the reproduction GRN [6]. Based on published data, regulatory relationship between fifteen ABC and key non-ABC genes were translated into a discrete Boolean network model. The state of each gene is updated according to the states of the genes that directly regulate it, via a set of logical rules derived from a survey of molecular genetic experimental data. Model simulations for all possible starting states showed that the network converges to a few steady states that correspond to expression patterns observed in each of the primordial cell types (i.e. inflorescence, sepals, petals, stamens or carpels) and are in agreement with the phenotypes predicted by the ABC genetic model for both wild type and mutants. Analysis of the simulation results and the logical rules derived from published data led the authors to speculate that the MADS-domain transcription factor AGAMOUS (AG) is involved in a positive feedback loop to maintain its own expression; this prediction was experimentally confirmed by a later study [7] .

A quantitative model that captures not only the regulatory relationships but also the kinetics of MADS-domain complex formation was later proposed to represent a GRN for organ-fate determination in Arabidopsis [8]. For this work, transcriptional regulation of six genes representing the five gene classes in the ABCDE model were modelled using ordinary differential equations (ODEs). For each gene, there is an ODE describing how the gene expression level is influenced by the concentrations of its regulators. Based on gene expression data, whorl-specific concentrations were estimated, which were then used to estimate the various model parameters. These parameters describe the interaction affinity of the various MADS-domain protein dimers, the binding affinity of these dimers to target promoters and the decay of gene products into non-functional components. The model generates continuous time-course expressions for the involved genes in the different floral whorls that reasonably match experimental data. It has to be noted that such a model provides more detailed

information about the network's dynamics than the pioneering discrete-network model [6], but contains many more parameters which have to be fitted using experimental data.

ODEs are widely accepted as modelling framework for GRNs. However, stochastic framework has also been shown important specially when the number of molecules involved is small or the time scale is short, or both [9]. Based on this, Lenser et al. [10] proposed stochastic models for representing the regulatory relationships and interactions between non-MADS box class B genes from the ABC model. Three hypotheses for the regulatory mechanism were investigated by analysing three respective stochastic models for the autoregulatory upregulation loop between *DEF*-like and *GLO*-like genes. By comparing the models, the case of obligate heterodimerization between DEF-like and GLO-like proteins provided a more certain decision behaviour than the cases in which heterodimerization is only facultative. Overall, this result suggested that obligate heterodimerization evolved to confer robustness of cell-fate organ identity decisions in the presence of stochastic noise.

The dynamics of a regulatory network can also be analysed using Petri net [11,12] models. These have the advantage of being accessible to biologists, duo to an intuitively understandable graphical notation. A Petri net is represented by a network model with two different types of nodes: *places* and *transitions*. *Places* represent components of the system (e.g. mRNA, protein), while *transitions* correspond to events that can change the state of the resources (e.g. degradation, transcription, translation, binding, transport). Weighted arcs (directed edges) connect *places* with *transitions*, thus depicting the regulatory relationships (e.g. activation, repression.) between resources and events. The state of a Petri net at each time is represented by the number of *tokens* associated to each *place*; which are dynamically controlled by the rules associated to the processes represented by the edges. Kaufmann et al. [13] presented a model based on these Petri nets to study the molecular interactions in a flower development network . That model included both direct regulation between genes (mediated by proteins) and the formation and regulatory effect of heterodimeric transcription factor complexes, thus allowing the simulation of the floral quartet concentration dynamics. This model showed that complex formation attenuates stochastic fluctuations in gene expression thus enabling more robust organ-specific expression patterns.

With respect to the modelling of GRNs involved in flowering time control, some of the modelling approaches mentioned above, such as Boolean networks or Petri Nets, in which time is not explicitly present, are less suitable. An ODE approach can serve as a framework for modelling the GRN underlying flowering time control. In particular, the control of flowering by photoperiod depends on an integration of external signals captured by photoreceptors and endogenous rhythm controlled by the circadian clock [14,15]. This makes mathematical models for the Arabidopsis clock gene circuit, as recently reviewed, very relevant [16] . One important aspect is the regulation of *FT,* which is a major direct target of the clock-regulated gene *CONSTANS* (*CO)* [17]. Salazar et al. [18] used an ODE-based

modelling to reconstruct the rhythmic regulation of *CO* by the circadian clock, and the subsequent effect of *CO* expression on the regulation of *FT*. To account for the effect of photoperiod on gene expression, the model assumed an explicit role of post-translational regulation: CO protein is stabilized during the daytime, but rapidly degrades during the night. Thus only the peak of *CO* mRNA that occurs in the light leads to CO protein accumulation and, therefore, *FT* activation. Interestingly, the expression levels of *FT* were simulated for different photoperiod cycles (light/dark), which indicated a non-linear relationship between observed flowering time and the amount of *FT* transcribed over a cycle-period. In the leaves, *FT* is not only regulated by the photoperiod sensors but also by temperature-related cues [15] via an *FLOWERING LOCUS C* (*FLC*)-mediated mechanism. Recently, the regulatory relationship between *FLC* and *FT* under the influence of temperature and photoperiod was modelled for the perennial *Arabidopsis halleri* [19]. Explicit information about temperature and photoperiod was used to simulate ODEs describing gene expression levels. Assessment of gene expression simulated for different temperatures allowed changes in the perennial flowering cycle to be forecasted under a climate change scenario. Finally, a recent ODE model for the transcriptional regulation of five key integrators of flowering time [20] was used to explore mechanistically how different feedback loops affect flowering time. For this work, each of the five components of the network is represented by a "hub". The idea is that a hub represents a set of genes and proteins that contribute to the same function. For instance, the AP1 hub represents the key integrators that determine the timing of floral transition; while FT hub represents the activity of at least FT and its interacting homolog TWIN SISTER OF FT (TSF). Interestingly, simulated AP1 hub activity together with the measured rosette and cauline leaf numbers were used to time the key steps of the floral transition, such as the switch from rosette to cauline leaf production or the end of flowering. This is important because it showed that flowering time can be predicted on the basis of e.g. *AP1*expression. Based on that, the flowering landscape was studied by simultaneously varying levels of two hubs, FT and TERMINAL FLOWER1 (TFL1). This enabled the authors to study closely the balance the balance of the FT and TFL1 hubs and how this correlates with the flowering behaviour. These examples show that GRN modelling enables various qualitative and quantitative predictions on network output and demonstrate the importance of such modelling approaches in understanding how the complex GRNs in plant reproduction fulfil their function. The use of mathematical modelling will facilitate a better insight into GRN complexity, allowing the consequences of network perturbations to be predicted. Current models are primarily based on gene expression profiles, while information about non-coding RNAs, protein-DNA and protein-protein interactions is increasingly produced and becoming incorporated into GRNs. Nevertheless, the experimentally validated protein-protein interactome is still quite sparse [21,22] and therefore we still rely on predicting the Arabidopsis interactome using orthology relationships, gene ontology and co-expression [23-25].

Multidimensional data analysis for understanding GRNs underlying plant reproduction

After the completion of the *Arabidopsis thaliana* genome [26], with additional plant genomes sequenced [27] or in progress, there has emerged a need to develop strategies to connect the various 'omics' results in order to address major questions in plant sciences. As noted by [2], enormous progress can be made when discovering emergent properties that connect the multiple dimensions of the data and thus bridging the different levels of systems organization. To give one example taken from [2], a current plant science challenge is to unravel how plants manage growth and development in response to biotic and abiotic stresses. This requires detailed understanding of plant functioning that tightly links molecular signalling networks to plant–community–abiotic environment networks. Another example also from [2], plant breeders need to increase agricultural productivity while decreasing the ecological footprint. This requires a holistic understanding that couples multiple levels of systems organizations considering for instance how sequence variation affects biological systems from cells to communities. In order to unveil the properties that bridge the different levels of systems organization, some researchers emphasize the role of modelling, whereas others stress multidimensional data analysis [3]. In the previous section I discussed application of mathematical modelling for plant reproduction networks. In this section I will introduce multidimensional data analysis (i.e. multi-'omics' data integration and analysis [28]) towards understanding GRNs underlying plant reproduction. It is focussed on protein-DNA binding, protein-protein interaction, gene and protein expression, sequence variation and phenotypic data and the integration of these types of data. The focus on these features is because they have been the most instrumental to unveil properties of the GRNs that control the major steps in plant reproduction.

TF binding and gene expression

Successful examples of combining 'omics' technologies are found in studies that integrate genome-wide TF DNA-binding profiles with gene expression studies. Some of these example are further detailed below. Experimentally, a current challenge is to unravel the aspects of spatial, temporal and combinatorial gene expression and regulation in the current networks; whilst from a bioinformatics and biostatistics perspective, the challenge is the development of tools, pipelines and analytical methods that will process these two 'omics' data. This integration could lead to: 1) infer the mode of action of the TFs (activator or repressor); 2) to determine the targets of the TFs; and 3) to elucidate and analyse the importance of motifs recognized by TFs. Software packages [29,30], webservers [31], pipelines [32], scripts [33,34] and R-packages [35] to analyse protein-DNA binding data integrated with gene expression data have recently been developed.

Recent studies of TF DNA-binding profiles combined with gene expression analyses have shown that there is only a weak correlation between binding of a TF and changes in expression of its target genes

[36]. An explanation could be that multiple TF binding events or co-factors are needed for gene regulation. In such a scenario only a specific combination of TF binding events will trigger changes in expression. Experimentally, sequential ChIP analysis [37] could be used to identify TF co-binding and may result in a better insight into the regulation of gene expression. Another explanation could be that a binding event to a single *cis*-regulatory element is not sufficient to drive expression, while binding of a TF to multiple sites, allowing a conformational change of the DNA, is needed to regulate gene expression. In this case, new techniques, such as chromatin capture [38] and ChIA-Pet [39,40], could be used to characterise *cis*-regulatory element interactions and their role in gene regulation. Studies that combine TF DNA binding profiling techiniques with gene expression analysis can lead to hypothesis about the molecular mechanisms underlying gene regulation.

A genome-wide TF DNA binding profiling study in combination with transcriptome analysis [41] showed that the flowering orchestrator *AP1* acts as both an activator and repressor of transcriptional regulation depending on the precise stage of flower development, indicating a dynamic mode of action. In a similar observation, the floral organ identity genes *AP3* and *PI* were also shown to act as both activators and repressors. Furthermore, it was suggested that their transcriptional roles are likely to be determined by the composition of the TF complexes [42]. The same was observed and suggested for *APETALA2 (AP2)* [30]. It may therefore be expected that from such combination of 'omics' approaches more TFs that are generally considered to be solely activators or repressors appear to have multiple mode of actions on the transcription of a diverse set of target genes. As demonstrated by these examples, a first step to obtain mechanistic insights in gene regulation is the integration of protein-DNA binding profiling with transcriptomics. When we are able to differentiate the down-regulated direct targets from the up-regulated targets, we can examine closely the extent to which either DNA binding sequences or combinatorial patterns determine the role of the TF as activator or repressor. A next step would be to interrogate protein-protein interaction networks to examine the binding specificities of the interacting partners of a given TF.

TF binding and protein-protein interaction data

As previously mentioned, protein-protein interaction studies will be required to enable us to understand TF-DNA interaction specificity. Thus, the value of genome-wide interactome information for understanding the combinatorial mode of action of TFs is unquestinable. A recent genome-wide interactome [43] analysis has elucidated the composition of several protein-protein complexes. In addition to that, protein family-based interaction networks [44,45] have elucidated the composition of many MADS-box TFs known to be involved in plant reproduction. However, the techniques currently used to generate these large-scale protein-protein interaction maps, such as yeast-two-hybrid assay, often suffer from an inflated rate of false positives [46]. For this reason, experimental validations of interactions detected by large-scale interaction screening are often required. Alternatively, protein-

protein interactions can be predicted [47]. The main bioinformatics and biostatistics challenge is to analyse the protein-protein interaction data, with all its inherent limitation, together with data from multiple TFs DNA binding profiling studies. This because protein-protein interactions are one of the key determinants of DNA binding specificity for many relevant TFs. Advanced proteomics [48] approaches combined with transcriptomics could be used to systematically study protein interactions within a regulatory network in order to e.g. detect post- transcriptional modifications or characterizing TF complexes [22]. SELEX (Systematic Evolution of Ligands by Exponential Enrichment) is a powerful method to characterise TF binding sites where different TF dimers can be tested for their DNA binding specificity. This has been applied to factors in the flowering GRN. For example, Moyroud et al. [49] applied SELEX coupled to Next Generation Sequencing (NGS) to determine the preferred binding sites of LFY.

Since several aspects of protein-DNA and protein-protein binding specificities are encoded in the protein sequences, another bioinformatics challenge is to catalogue, at a network-wide scale, the functional parts of the protein, such as the protein-DNA and protein-protein binding domains. This could be achieved by integrating sequence, structure and protein-protein interaction information. Such catalogue can aid experimentalists to e.g. study the effect of small protein mutations on intermediate phenotypes.

Integration of sequence variation

Recent genetic studies started addressing how sequence variation reflects on differences in the molecular mechanisms underlying adaptive traits, such as flowering time control. For example, one study that combines gene expression data with phenotypic and sequence variation information of 192 *Arabidopsis* accessions [50] has revealed that loss-of-function mutations in the coding region of the flowering time gene *FRIGIDA* can confer a strong selective advantage. Similarly, a more recent study used a similar approach with 16 *Arabidopsis* accessions [51] to show that *cis*-regulatory mutations in the promoter of *CO* are associated with variation in gene expression, and consequently with variations in phenotype, demonstrating that *cis*-regulatory mutations also play a strong role in the evolution of flowering time. These studies focused on associations of a single gene. At a larger-scale, the computational challenge is to harness the data being generated by projects such as the 1001 *Arabidopsis* genomes project [52] to comprehensively catalogue the variants that are associated with adaptive phenotypic traits. In this direction, studies that combine genomics with phenomics such as genome-wide association studies (GWAS) have been performed to quantify the association of polymorphisms to phenotypic traits, including flowering time [53].

In spite of the advances, a major challenge in the interpretation of GWAS results for cataloguing and understanding the role of polymorphisms comes from the fact that the detected associations point to

relatively large regions of correlated variants (linkage disequilibrium region). This makes it difficult to precisely identify the single nucleotide polymorphism (SNP) that has a biological link with the phenotype [54]. In addition, the genetic backgrounds of the population often produce a confounding effect due to the presence of a structure in the population, which results in a bias in the statistical analysis [55] .This inflates the numbers of false positive associations. Therefore, GWAS can be seen as starting points towards understanding the role of polymorphisms on the adaptive phenotypic trait, while often other 'omics' data are integrated with the goal to validate, strengthen and refine the GWAS signals [56]. This can be done e.g. by 1) incorporating gene expression data to augment the GWAS signal and increase its power [57]; or to refine GWAS results by 2) using information about protein-DNA binding profiles in order to filter those associations that overlap the position of known regulatory elements [54], or by 3) using information about evolutionarily constrained regions in coding sequences (such as the protein-DNA or protein-protein domains) [58] in order to filter those associations with an effect on protein function.

As genetic variation studies are combined with other "omics" data, such as phenomics, gene expression, protein-protein interaction (PPI) data and protein-DNA binding profiling studies, we will be able to comprehensively catalogue the variants that have an effect on specific phenotypes. Once important polymorphisms have been catalogued, the next step is to understand their roles. In this respect, it has been shown that variations in several distinct genes can result in similar phenotypes (heterogeneity), whilst some phenotypes may only be observed when a specific combination of perturbations occurs (epistasis, redundancy) [59]. Therefore, in order to understand the molecular mechanisms that underlie the genetic variations, it is very relevant to study the genetic variations in a network context. In this regard, recent studies have proposed the integration of GWAS results with PPI networks. For example, it has been proposed by Xu et al. [60] in human studies that the associations of genes and diseases can be revealed by analysing topological features of PPI. For that, five measures - which also include the GWAS gene-disease association, computed from diverse combinations of sub-networks were integrated to determine the likelihood of the genes being associated to a disease. More recently, Han et al. [61] successfully made use of an R package [62] that screens the PPI network for enriched sub-networks whose genes show low *p*-value in GWAS datasets. These studies have the advantage of 1) possibility to identify genes that could be missed by traditional approaches that search SNP in an one-gene-to-phenotype manner; and of 2) unveiling sub-networks of interacting genes linked to the phenotype. Both PPI-based approaches offer ways to identify SNPs that co-occur in the encoded interacting proteins, but they do not specify the molecular mechanisms that underlie the variation, i.e. it is not clear if the SNP disrupts the protein-protein interactions. To determine that, one could use, depending on the availability of information, either annotations and predictions of locations of functional parts of the protein, or methods that predict the effect of SNPs on PPI given the structures of the proteins [63]. In a broader strategy than solely using PPI networks,

pathway-based analyses have been proposed to screen for associations of SNPs in genes that play a role in the pathway linked to the phenotype [64,65]. On the one hand, pathway-based approaches have the advantage of providing more biological insight into the mechanisms underlying the mutations, however, on the other hand, when applied to complex phenotypic traits such as flowering time control these approaches have the disadvantage that they cannot detect SNPs in genes that work across pathways. To overcome that, a next step would be to integrate approaches that identify relevant polymorphisms within a GRN. This could be achieved, as recently envisioned by [66], by firstly identifying the SNPs in a GRN context, and secondly by using GRN models to study the effect of polymorphisms on intermediate phenotypes, such as gene expression, protein-protein interaction or cooperativity profiles and ultimately, the subsequent effect on the trait phenotype.

In conclusion, multi-'omics' data integration and analysis must be addressed to take full advantage of plant systems biology towards unveiling and understanding molecular mechanisms underlying GRNs. This includes not only integrating data of the same level of systems organization (e.g. transcriptomics with proteomics) but also studying the interplay between the levels; for example by combining evolutionary and developmental aspects to study molecular mechanisms of genetic variation among individuals within species, or variation between related species.

Scope and outline of this thesis

The research described in this thesis aimed to integrate several X-omics data to obtain information about the functioning of a biological system and the underlying molecular mechanisms. It is focused on three main key objectives: (1) understanding the relationship between expression patterns and a phenotypic trait by modelling the GRN of flowering time control, (2) understanding how SNPs in *cis*-elements can explain the existing genetic diversity in Arabidopsis accessions related to flowering time control, (3) developing a framework for predicting protein-protein binding motifs.

In Chapter 2, we modelled with ordinary differential equations (ODEs) the regulatory network constituted by a set of core genes controlling Arabidopsis flowering time in order to quantitatively analyse the relationship between their expression levels and the flowering time response. We considered a core gene regulatory network composed of *FLOWERING LOCUS T (FT)* and seven transcription factor genes: *SHORT VEGETATIVE PHASE (SVP), FLOWERING LOCUS C (FLC), AGAMOUS-LIKE 24 (AGL24), SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 (SOC1), APETALA1 (AP1), LEAFY (LFY)* and *FLOWERING LOCUS D (FD).*

Chapter 3 presents an approach to identify single nucleotide polymorphisms (SNPs) that may have a role in the regulation of Arabidopsis flowering time genes. Reliable identification of genetic variants that affect gene regulation is still a challenge in systems biology and is expected to play an important role in the molecular characterization of complex traits [67]. For this chapter we used data from ChIP-

Seq experiments of transcription factors involved in the control of flowering time to identify variants in a diverse set of Arabidopsis accessions with a possible effect on flowering time.

In Chapter 4, we present the results of genome-wide prediction of protein-protein binding sites form the Arabidopsis interactome. We developed a variant of the SLIDER method, that uses a protein interaction network to locate binding sites in the sequence of interacting proteins [68,69]; we modified the algorithm to allow various types of biological knowledge into account. In addition, we parameterized the method to predict motifs from the available Arabidopsis interactome [43]. This new method is named SLIDERBio [70] and is available for download at http://www.ab.wur.nl/sliderbio. We interrogated the interactome data to formulate testable hypotheses for the molecular mechanisms affected by protein sequence mutations. Examples include proteins relevant for various developmental processes, including flowering.

Finally, Chapter 5 discusses the future applications of models involving GRNs and the integration of other omics data, as well as lessons learned from the limitations of modelling frameworks for systems biology.

References

1. Blazquez MA, Ferrandiz C, Madueno F, Parcy F (2006) How floral meristems are built. Plant Mol Biol 60: 855-870.
2. Keurentjes JJ, Angenent GC, Dicke M, Dos Santos VA, Molenaar J, et al. (2011) Redefining plant systems biology: from cell to ecosystem. Trends Plant Sci 16: 183-190.
3. Yuan JS, Galbraith DW, Dai SY, Griffin P, Stewart CN, Jr. (2008) Plant systems biology comes of age. Trends Plant Sci 13: 165-171.
4. Smaczniak C, Immink RG, Angenent GC, Kaufmann K (2012) Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. Development 139: 3081-3098.
5. Kaufmann K, Pajoro A, Angenent GC (2010) Regulation of transcription in plants: mechanisms controlling developmental switches. Nat Rev Genet 11: 830-842.
6. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER (2004) A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. Plant Cell 16: 2923-2939.
7. Gomez-Mena C, de Folter S, Costa MM, Angenent GC, Sablowski R (2005) Transcriptional program controlled by the floral homeotic gene AGAMOUS during early organogenesis. Development 132: 429-438.
8. van Mourik S, van Dijk AD, de Gee M, Immink RG, Kaufmann K, et al. (2010) Continuous-time modeling of cell fate determination in Arabidopsis flowers. BMC Syst Biol 4: 101.
9. Shmulevich I, Aitchison JD (2009) Deterministic and stochastic models of genetic regulatory networks. Methods Enzymol 467: 335-356.
10. Lenser T, Theissen G, Dittrich P (2009) Developmental robustness by obligate interaction of class B floral homeotic genes and proteins. PLoS Comput Biol 5: e1000264.
11. Karlebach G, Shamir R (2008) Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol 9: 770-780.
12. Chaouiya C (2007) Petri net modelling of biological networks. Brief Bioinform 8: 210-219.
13. Kaufmann K, Nagasaki M, Jauregui R (2011) Modelling the Molecular Interactions in the Flower Developmental Network of Arabidopsis thaliana. Stud Health Technol Inform 162: 279-297.
14. Hayama R, Coupland G (2003) Shedding light on the circadian clock and the photoperiodic control of flowering. Curr Opin Plant Biol 6: 13-19.
15. Song YH, Ito S, Imaizumi T (2013) Flowering time regulation: photoperiod- and temperature-sensing in leaves. Trends Plant Sci 18: 575-583.
16. Bujdoso N, Davis SJ (2013) Mathematical modeling of an oscillating gene circuit to unravel the circadian clock network of Arabidopsis thaliana. Front Plant Sci 4: 3.
17. Yanovsky MJ, Kay SA (2002) Molecular basis of seasonal time measurement in Arabidopsis. Nature 419: 308-312.
18. Salazar JD, Saithong T, Brown PE, Foreman J, Locke JC, et al. (2009) Prediction of photoperiodic regulators from quantitative gene circuit models. Cell 139: 1170-1179.
19. Satake A, Kawagoe T, Saburi Y, Chiba Y, Sakurai G, et al. (2013) Forecasting flowering phenology under climate warming by modelling the regulatory dynamics of flowering-time genes. Nat Commun 4: 2303.
20. Jaeger KE, Pullen N, Lamzin S, Morris RJ, Wigge PA (2013) Interlocking feedback loops govern the dynamic behavior of the floral transition in Arabidopsis. Plant Cell 25: 820-833.
21. Chatr-Aryamontri A, Breitkreutz BJ, Heinicke S, Boucher L, Winter A, et al. (2013) The BioGRID interaction database: 2013 update. Nucleic Acids Res 41: D816-823.
22. Smaczniak C, Immink RG, Muino JM, Blanvillain R, Busscher M, et al. (2012) Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. Proc Natl Acad Sci U S A 109: 1560-1565.

23. De Bodt S, Proost S, Vandepoele K, Rouze P, Van de Peer Y (2009) Predicting protein-protein interactions in Arabidopsis thaliana through integration of orthology, gene ontology and co-expression. BMC Genomics 10: 288.
24. Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY (2010) Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. Nat Biotechnol 28: 149-156.
25. Lin M, Shen X, Chen X (2011) PAIR: the predicted Arabidopsis interactome resource. Nucleic Acids Res 39: D1134-1140.
26. Arabidopsis Genome I (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796-815.
27. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, et al. (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40: D1178-1186.
28. Joyce AR, Palsson BO (2006) The model organism as a system: integrating 'omics' data sets. Nat Rev Mol Cell Biol 7: 198-210.
29. Wang S, Sun H, Ma J, Zang C, Wang C, et al. (2013) Target analysis by integration of transcriptome and ChIP-seq data with BETA. Nat Protoc 8: 2502-2515.
30. Maienschein-Cline M, Zhou J, White KP, Sciammas R, Dinner AR (2012) Discovering transcription factor regulatory targets using gene expression and binding data. Bioinformatics 28: 206-213.
31. Qin J, Li MJ, Wang P, Zhang MQ, Wang J (2011) ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. Nucleic Acids Res 39: W430-436.
32. Franco-Zorrilla JM, Lopez-Vidriero I, Carrasco JL, Godoy M, Vera P, et al. (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc Natl Acad Sci U S A 111: 2367-2372.
33. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, et al. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38: 576-589.
34. Jeffery IB, Madden SF, McGettigan PA, Perriere G, Culhane AC, et al. (2007) Integrating transcription factor binding site information with gene expression datasets. Bioinformatics 23: 298-305.
35. Zhu LJ, Gazin C, Lawson ND, Pages H, Lin SM, et al. (2010) ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. BMC Bioinformatics 11: 237.
36. O'Maoileidigh DS, Graciet E, Wellmer F (2014) Gene networks controlling Arabidopsis thaliana flower development. New Phytol 201: 16-30.
37. Oh E, Zhu JY, Wang ZY (2012) Interaction between BZR1 and PIF4 integrates brassinosteroid and environmental responses. Nat Cell Biol 14: 802-809.
38. Stadhouders R, Kolovos P, Brouwer R, Zuin J, van den Heuvel A, et al. (2013) Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. Nat Protoc 8: 509-524.
39. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, et al. (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462: 58-64.
40. Zhang J, Poh HM, Peh SQ, Sia YY, Li G, et al. (2012) ChIA-PET analysis of transcriptional chromatin interactions. Methods 58: 289-299.
41. Kaufmann K, Wellmer F, Muino JM, Ferrier T, Wuest SE, et al. (2010) Orchestration of floral initiation by APETALA1. Science 328: 85-89.
42. Wuest SE, O'Maoileidigh DS, Rae L, Kwasniewska K, Raganelli A, et al. (2012) Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA. Proc Natl Acad Sci U S A 109: 13452-13457.
43. Arabidopsis Interactome Mapping C (2011) Evidence for network evolution in an Arabidopsis interactome map. Science 333: 601-607.
44. Immink RG, Tonaco IA, de Folter S, Shchennikova A, van Dijk AD, et al. (2009) SEPALLATA3: the 'glue' for MADS box transcription factor complex formation. Genome Biol 10: R24.

45. de Folter S, Immink RG, Kieffer M, Parenicova L, Henz SR, et al. (2005) Comprehensive interaction map of the Arabidopsis MADS Box transcription factors. Plant Cell 17: 1424-1433.
46. Morsy M, Gouthu S, Orchard S, Thorneycroft D, Harper JF, et al. (2008) Charting plant interactomes: possibilities and challenges. Trends Plant Sci 13: 183-191.
47. van Dijk AD, ter Braak CJ, Immink RG, Angenent GC, van Ham RC (2008) Predicting and understanding transcription factor interactions based on sequence level determinants of combinatorial control. Bioinformatics 24: 26-33.
48. Smaczniak C, Li N, Boeren S, America T, van Dongen W, et al. (2012) Proteomics-based identification of low-abundance signaling and regulatory protein complexes in native plant tissues. Nat Protoc 7: 2144-2158.
49. Moyroud E, Minguet EG, Ott F, Yant L, Pose D, et al. (2011) Prediction of regulatory interactions from genome sequences using a biophysical model for the Arabidopsis LEAFY transcription factor. Plant Cell 23: 1293-1306.
50. Shindo C, Aranzana MJ, Lister C, Baxter C, Nicholls C, et al. (2005) Role of FRIGIDA and FLOWERING LOCUS C in determining variation in flowering time of Arabidopsis. Plant Physiol 138: 1163-1173.
51. Rosas U, Mei Y, Xie Q, Banta JA, Zhou RW, et al. (2014) Variation in Arabidopsis flowering time associated with cis-regulatory variation in CONSTANS. Nat Commun 5: 3651.
52. Weigel D, Mott R (2009) The 1001 genomes project for Arabidopsis thaliana. Genome Biol 10: 107.
53. Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, et al. (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465: 627-631.
54. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M (2012) Linking disease associations with regulatory information in the human genome. Genome Res 22: 1748-1759.
55. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am J Hum Genet 67: 170-181.
56. Ioannidis JP, Thomas G, Daly MJ (2009) Validating, augmenting and refining genome-wide association signals. Nat Rev Genet 10: 318-329.
57. Li L, Kabesch M, Bouzigon E, Demenais F, Farrall M, et al. (2013) Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. Front Genet 4: 103.
58. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248-249.
59. Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, et al. (2014) Genomics and the origin of species. Nat Rev Genet 15: 176-192.
60. Xu J, Li Y (2006) Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics 22: 2800-2805.
61. Han S, Yang BZ, Kranzler HR, Liu X, Zhao H, et al. (2013) Integrating GWASs and human protein interaction networks identifies a gene subnetwork underlying alcohol dependence. Am J Hum Genet 93: 1027-1034.
62. Jia P, Zheng S, Long J, Zheng W, Zhao Z (2011) dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. Bioinformatics 27: 95-102.
63. Zhao N, Han JG, Shyu CR, Korkin D (2014) Determining Effects of Non-synonymous SNPs on Protein-Protein Interactions using Supervised and Semi-supervised Learning. PLoS Comput Biol 10: e1003592.
64. Wang K, Zhang H, Kugathasan S, Annese V, Bradfield JP, et al. (2009) Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. Am J Hum Genet 84: 399-405.
65. Wang K, Li M, Hakonarson H (2010) Analysing biological pathways in genome-wide association studies. Nat Rev Genet 11: 843-854.
66. Marjoram P, Zubair A, Nuzhdin SV (2014) Post-GWAS: where next? More samples, more SNPs or more biology? Heredity (Edinb) 112: 79-88.

67. Hudson TJ (2003) Wanted: regulatory SNPs. Nat Genet 33: 439-440.
68. Boyen Peter NF, Van Dyck Dries. (2010) Mining Correlated Motifs in Protein-Protein Interaction Networks. . ICDM '09: Proceedings of the 2009 Ninth IEEE International Conference on Data Mining: 716-721.
69. Boyen P, Neven F, van Dyck D, Valentim FL, van Dijk AD (2013) Mining minimal motif pair sets maximally covering interactions in a protein-protein interaction network. IEEE/ACM Trans Comput Biol Bioinform 10: 73-86.
70. Leal Valentim F, Neven F, Boyen P, van Dijk AD (2012) Interactome-wide prediction of protein-protein binding sites reveals effects of protein sequence variation in Arabidopsis thaliana. PLoS One 7: e47022.

Chapter 2

A quantitative and dynamic model of the Arabidopsis flowering time gene regulatory network

Felipe Leal Valentim[1], S. van Mourik[2,6], D. Posé[3,$], M.C. Kim[3,§], M. Schmid[3], R.C.H.J. van Ham[4], M. Busscher[1], G.F. Sanchez-Perez[1,7], J. Molenaar[2], G.C. Angenent[1,5], R.G.H. Immink[1], A. D. J. van Dijk[1,2,6]

[1] Bioscience, Plant Research International, Bioscience, Wageningen, The Netherlands.

[2] Biometris, Wageningen UR, The Netherlands

[3] Max Planck Institute for Developmental Biology, Molecular Biology, Tübingen, Germany

[4] Keygene N.V., Wageningen, The Netherlands

[5] Laboratory of Molecular Biology, Wageningen University, Wageningen, The Netherlands

[6] Netherlands Consortium for Systems Biology, Amsterdam

[7] Chair group Bioinformatics, Wageningen University, Wageningen, The Netherlands

[$] Current Address: Instituto de Hortofruticultura Subtropical y Mediterránea, Universidad de Málaga–Consejo Superior de Investigaciones Científicas, Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, 29071 Málaga, Spain

[§] Current Address: Division of Applied Life Science (BK21 Plus), Gyeongsang National University, Jinju, Korea

PLoS ONE, *in press*

Abstract

Various environmental signals integrate into a network of floral regulatory genes leading to the final decision on when to flower. Although a wealth of qualitative knowledge is available on how flowering time genes regulate each other, only a few studies incorporated this knowledge into predictive models. Such models are invaluable as they enable to investigate how various types of inputs are combined to give a quantitative readout. To investigate the effect of gene expression disturbances on flowering time, we developed a dynamic model for the regulation of flowering time in *Arabidopsis thaliana*. Model parameters were estimated based on expression time-courses for relevant genes, and a consistent set of flowering times for plants of various genetic backgrounds. Validation was performed by predicting changes in expression level in mutant backgrounds and comparing these predictions with independent expression data, and by comparison of predicted and experimental flowering times for several double mutants.

Remarkably, the model predicts that a disturbance in a particular gene has not necessarily the largest impact on directly connected genes. For example, the model predicts that *SUPPRESSOR OF OVEREXPRESSION OF CONSTANS* (*SOC1*) mutation has a larger impact on *APETALA1 (AP1)*, which is not directly regulated by SOC1, compared to its effect on *LEAFY* (*LFY*) which is under direct control of SOC1. This was confirmed by expression data. Another model prediction involves the importance of cooperativity in the regulation of *APETALA1* (*AP1*) by LFY, a prediction supported by experimental evidence. Concluding, our model for flowering time gene regulation enables to address how different quantitative inputs are combined into one quantitative output, flowering time.

Introduction

Flowering at the right moment is crucial for the reproductive success of flowering plants. Hence, plants have evolved genetic and molecular networks integrating various environmental cues with endogenous signals in order to flower under optimal conditions [1]. Various input signals are received and transmitted by signal transduction pathways including the photoperiod pathway, the vernalization pathway, the ambient temperature pathway and the autonomous pathway [2]. Finally, the input from these pathways is integrated by a core set of flowering time integration genes ("integration network"). This regulation contributes to the adaptation of plants to different environmental conditions and facilitated the successful dispersion of flowering plants over the world [2].

The complexity of flowering time regulation is enormous, even when focusing on the network involved in integrating the various signals. To understand how gene disturbances influence flowering time, it is not only important to know which genes regulate each other, but also how strongly these genes influence each other. Hence, quantitative aspects of flowering time changes upon perturbations of input signals cannot be understood by merely assessing qualitatively which interactions are present.

To this end, a quantitative model describing how different genes in the network regulate each other is needed. Indeed, other complex plant developmental processes have been subject to extensive modeling efforts [3]. This includes processes such as the circadian clock [4-7], auxin signalling [8-11], photoperiod regulation of flowering time genes [12,13] and the development of floral organs [14-17], which all have been investigated in detail by computational models. These models enable to formalize biological knowledge and hypotheses, and, importantly, to investigate how various types of inputs are combined to give a quantitative readout.

Flowering time regulation has been extensively studied experimentally in the plant model species *Arabidopsis thaliana*. Substantial qualitative information is available about the factors involved and how these interact genetically. However, the information that is needed for quantitative and dynamic modelling is missing to a large extent. This includes comprehensive and standardized quantitative data on flowering time under various conditions and in different genetic backgrounds [18], and time series of expression for key flowering time integration genes in those backgrounds. In line with the scarcity of quantitative information useful for modelling, the floral transition in *Arabidopsis thaliana* has been scarcely studied using modeling approaches. Recently a few promising mathematical modeling approaches appeared aimed at modeling the floral transition in various plant species [19-21]. Dong et al. modeled a network of four genes involved in the floral transition in maize [19], and Satake et al. modeled a two-gene network in *Arabidopsis halleri* [21]. Earlier work on modelling *Arabidopsis thaliana* flowering time did not take genetic regulation into account or used a mainly qualitative approach [22]. Only very recently a first quantitative model of the *Arabidopsis thaliana* flowering time integration network was presented [20].

We aimed to obtain a mechanistic understanding of the *Arabidopsis thaliana* flowering time integration network, by investigating a core gene regulatory network composed of eight genes (Fig. 1): *SHORT VEGETATIVE PHASE (SVP), FLOWERING LOCUS C (FLC), AGAMOUS-LIKE 24 (AGL24), SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 (SOC1), APETALA1 (AP1), FLOWERING LOCUS T (FT), LEAFY (LFY)* and *FD*. Although certainly more genes are involved in integrating the various signals influencing the timing of the floral transition [2,23], we focused on these genes because a) we aim to model the core of the network responsible for flowering time regulation; and b) for these genes, the available experimental data renders a clear picture of their mutual interactions (see above; SI Table 1; Figure 1). In the leaves, SVP and FLC repress the transcription of *FT* [24-27]. FT is produced in the leaves and moves to the shoot apical meristem (SAM) [28,29]. FT has the potential to interact with FD [30,31] and complex formation is supposed to occur at the SAM, leading to activation of *SOC1* [32] and *AP1* expression [30,33]. *FLC* and *SVP* are also expressed in the SAM, where they repress the expression of *SOC1* [34-36]. *SOC1* integrates signals from multiple pathways and transmits the outcome to *LFY* [37,38], which is supposed to act at least partially via a positive feed-back loop in which *AGL24* is involved upon dimerizing with SOC1

[39]. In turn, LFY is a positive regulator of *AP1* [40] and of FD [20]. The commitment to flower is ascertained by a direct positive feed-back interaction between *AP1* and *LFY*. Once the expression of *AP1* is initiated, this transcription factor orchestrates the floral transition by specifying floral meristem identity and regulating the expression of genes involved in flower development [41]. Importantly, in comparison with the recently presented model of the floral transition in Arabidopsis [20] we included the key floral integrator genes *SOC1*, *SVP* and *AGL24* in our model.

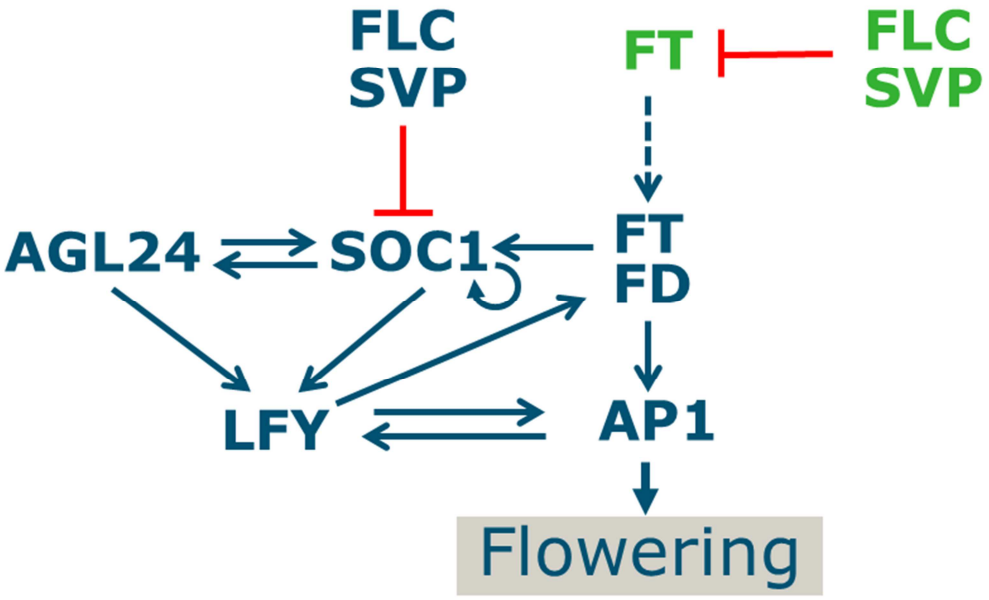


Figure 1: Network of flowering time integrator genes. Green indicates expression in leaf tissue, blue in meristem tissue. Red arrows represent repression, blue arrows activation. Dashed arrow represents FT transport. As indicated, *AP1* expression is used as a marker for the moment of the floral transition. This network was used to fit expression time-course data and to predict the effect of perturbations. Gene names are given in full in the text.

The above introduced interactions between the flowering time integration genes and the floral meristem identity genes at the end of the pathway allow to derive a set of Ordinary Differential Equations (ODEs) describing how genes in the network regulate each other. ODEs were chosen because they arise from continuum modelling of molecular interactions and allow quantitative analysis of the effect of perturbations on expression levels and finally on flowering time. Because of the above mentioned role of *AP1* as orchestrator of floral meristem identity specification, the moment at which the *AP1* expression level starts to rise is used as a proxy for flowering time in the model.

In order to build and validate an ODE model describing the network constituted by the eight selected genes, we obtained three quantitative datasets: i) gene expression time-courses of the selected eight genes in wild type; ii) flowering time of plants of different genetic backgrounds; and iii) expression data of the selected genes in the plants of those different genetic backgrounds. A key aspect of our

approach is that we estimate model parameters using the dynamic gene expression time-course data for the components of the model, in combination with flowering time data (datasets *i* and *ii*). We validated our model by comparing predicted expression time-courses for mutants in components of the network with experimental data (dataset *iii*). Finally, we obtained detailed understanding of how genes are affected by perturbation in other genes, via the regulatory interactions that constitute the network.

Results

Model building and parameter estimation

Given the importance of combining various input signals into a final decision to flower, a key question is how the integration network generates a quantitative response, i.e. how expression level perturbations of various magnitudes result in specified changes in other network components and finally in a change in flowering time. In order for the model to be able to link expression changes to changes in flowering time, we included *AP1*: expression of *AP1* indicates that the switch from vegetative to reproductive growth has occurred [42]. As such, we use the moment at which *AP1* expression rises above a certain threshold in our model as a proxy for the moment at which flowering starts (see Methods for details).

Our approach to investigate the network involves modelling by ordinary differential equations (ODEs), which describe how the expression level of each gene is influenced by the other genes. This regulation is described by Hill functions, which represent activation or repression by the various regulators. Genetic and molecular knowledge on the network structure is used as input to define those equations. Parameters in these equations represent interaction strengths and other biological or physical aspects of the system, and are estimated using wild-type gene expression time-course data. *FLC* and *SVP* are not known to be regulated by any of the genes included in our model, and for that reason, they are included as external input factors, that regulate one or more other genes in the model. In order to model FT transport to the shoot apical meristem [43], we assumed that the FT produced in the leaves reaches the meristem with a delay. An optimal parameter set, which includes the FT transport delay, was identified by fitting the equations to qRT–PCR time-course data from leaves and SAM-enriched material obtained from Arabidopsis plants grown at 23 ^{0}C under long-day (LD) conditions (Methods).

The genes in the core regulatory network of flowering time control cooperate to activate the flowering orchestrator *AP1* [41]. This allows proper timing of *AP1* expression and fine tuning of flowering time in response to different environmental cues. In wild type Arabidopsis, the *AP1* level remains barely detectable in the SAM until about day 13 after germination and then sharply increases (Fig. 2). As mentioned above, we use the moment at which *AP1* expression level rises as a marker to indicate that the transition to reproductive development is completed, which is interpreted as a predictor of

flowering time. Based on that, we developed a fitting strategy that besides of aiming at a good fit, optimizes the correlation between predicted and observed flowering time. To be able to do so, we obtained a consistent set of flowering time measurements for mutants of six of the genes in our network (SI Figure 1). Flowering time was measured as the number of rosette leaves (RL) present at flowering. To compare model predictions for flowering time, expressed in units of days, these were scaled to RL (Methods).

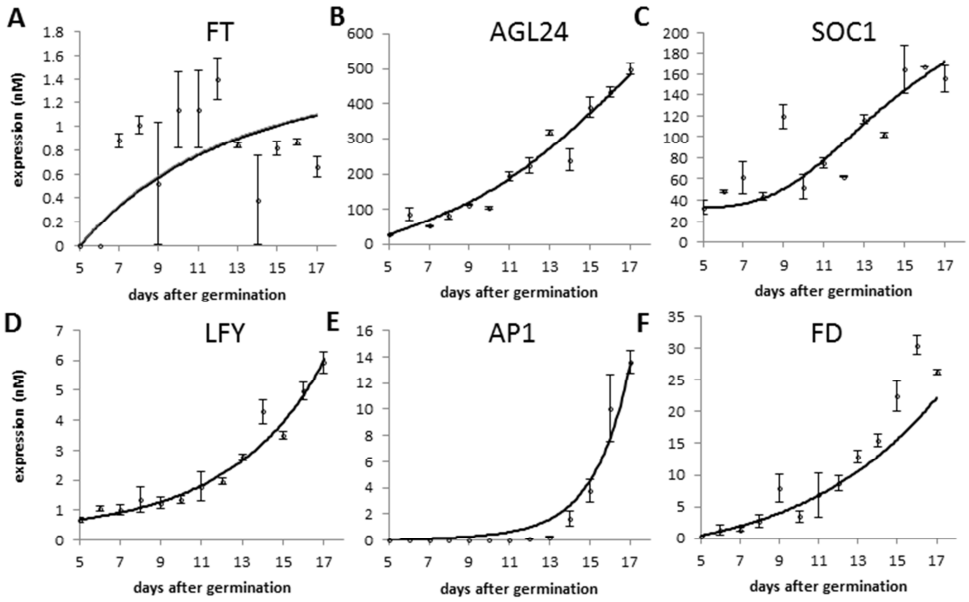


Figure 2: Experimental and simulated expression time-course of the genes in the integration network model. Gene expression was measured by qRT-PCR (shown as dots) of wild type plants grown under long-day conditions at 23°C (average and standard deviation are shown). The continuous lines show the simulated gene expression using the parameters estimated by data fitting. Note that *FLC* and *SVP* are not regulated by other components of the network and hence are present as input factors only, and their expression level is not simulated by the model. qRT-PCR data for *FT* was obtained from leaves; for the other genes, qRT-PCR data was obtained from meristem enriched material.

A total of 35 parameters in six equations were estimated from the time series data containing 13 data points (expression levels) per gene (SI Table 2-3; Figure 2). Given the variability in the data, the fit is satisfactory, as indicated by the value of the normalized root mean square error (nrmse). For *FT*, for which the data shows highest variability, the highest nrmse (27%) was obtained. For *SOC1*, the overall fit was good, but does not capture the data point at day 9, which deviates from the general trend in the time series, resulting in a nrmse of 19%. For *AP1* and *FD* the value of the nrmse was around 14%, and for *AGL24* and *LFY* it was 7%. The *FLC* and *SVP* expression data were used directly as input to the model; these are shown in SI Figure 2. Interestingly, for the data describing *AP1*, we could only obtain a good fit by introducing a particular value of one parameter describing how *AP1* is

regulated by LFY. As further discussed below, this parameter indicates DNA binding cooperativity for which indeed experimental evidence exists.

Simulated flowering times in various genetic backgrounds are shown in Figure 3A. There is one outlier in this plot (*ft-10*). Besides this exception, there is considerable agreement between data and predictions. Indeed, comparison of the Pearson correlation (R=0.85, including the outlier) with correlation obtained using randomized data demonstrates the significance of this result ($p<0.005$).

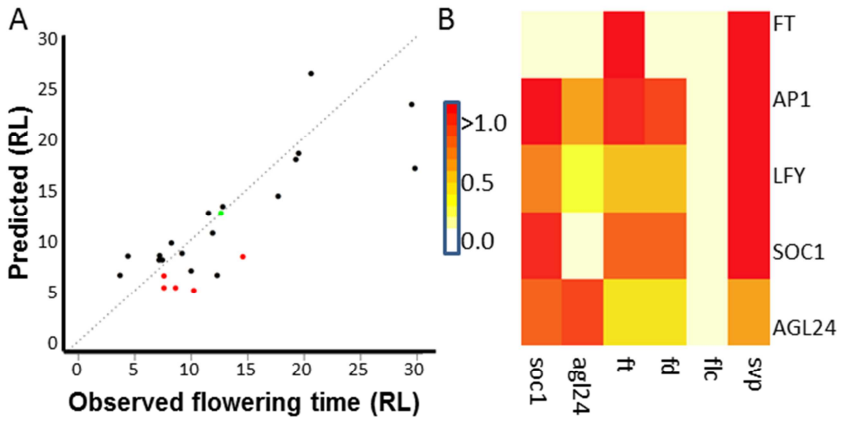


Figure 3: Model predictions and experiments in various mutant backgrounds. (A) Predicted vs. experimentally observed flowering time for mutants used in training the model (black) and for double mutants used for validation (red). Wild type flowering time is indicated in green. RL, rosette leaves: the more rosette leaves, the later flowering. **(B)** Prediction of expression changes; total change in expression over the simulated time-course is calculated, normalized against wild type; absolute value is reported to focus on the magnitude of the predicted expression change. Horizontal axis, mutants; vertical axis, genes for which expression change in mutant background is simulated. Note that FLC and SVP are not regulated by other genes in the model and hence, their expression level does not change upon any mutation. For comparison between predictions and experiments, see SI Figures 3 and 4.

Model validation

A key issue in our model is the mechanism by which the network is able to give a quantitative response to specific perturbations. How are changes in a given gene expression level transferred to other components of the network, and how does this impact flowering time? In order to validate model predictions of how changes in expression propagate through the network, we simulated the expression time-courses for mutants and obtained independent experimental data for comparison. For that, microarray experiments were used, which were carried out for wild type and four mutant backgrounds (*soc1*, *agl24*, *fd* and *flc*). In these experiments, a flowering inducing shift from short-day to long-day conditions was used [44]. As indicated by the value of Pearson's R (0.69; p-value = 0.003), the predicted expression level changes of flowering time genes upon upstream mutations show a

significant correlation with the experimental data (SI Fig. 3-4). Assessing the correlation per gene (across the different mutants) indicates similar correlation for each of the genes. However, assessing the correlation per mutation (across the different genes) indicates good predictive performance for *SOC1*, *FD* and *AGL24* mutations, but not for *FLC* mutation. The latter could be due to the low expression and limited role of *FLC* in the Col-0 background due to the *FRIGIDA (FRI)* mutation [45]. The comparison with the microarray dataset constitutes an independent evaluation of the predictive performance of the model, demonstrating that the model allows predicting the magnitude of the effect of a perturbation in one gene on other genes in a quantitative manner.

To further assess the predictive performance of the model, we analysed five double mutants in which over-expression of one gene was combined with knock-out of a second gene. In all cases, both genes involved activators of flowering (SI Fig. 1), implying that it is intuitively difficult to predict whether the double mutant will be early or late flowering. These mutants were not used in the parameter estimation stage. The resulting prediction performance was satisfactory (Fig. 3A): for four out of five cases, the prediction was qualitatively correct ("early flowering"). Quantitatively, the correlation between experimental and predicted flowering times was reasonable as well, although not significant at the p=0.05 level (Pearson R=0.75; p=0.1). It is good to realize that no perfect fit was expected in this case because of variable temporal and spatial overexpression levels due to the usage of the 35S promoter [46].

Spread of perturbations through the network

As a first example of quantitative understanding of flowering time regulation, we analysed the predicted expression changes in various mutant backgrounds (Fig. 3B). A key question here is how gene expression perturbations spread through the network. We found that the model predicts that the spread of a perturbation is not in all cases directly related to the position that different genes have in the network (Fig. 3B). For example, the effect of mutating *SOC1* on *LFY* is smaller than its effect on *AP1*, although SOC1 regulates *LFY* and does not directly regulate *AP1*, but only indirectly via LFY. Analysis of the regulatory interactions and the associated parameters in the model allows rationalizing such differences. For the above-mentioned different magnitudes of the effect of *soc1* mutation on *LFY* compared to its effect on *AP1*, it is relevant that the estimated expression activation strength (parameter β) for the influence of SOC1 on *LFY* (β_7) is much smaller than that for the influence of LFY on *AP1* (β_9; SI Table 3). This means that the model predicts that a change in SOC1 will give rise to a relatively small change in *LFY*, which however will be amplified by LFY regulating *AP1*. This effect is visible in the experimental microarray data as well, where in the *soc1* mutant background *LFY* expression is much less affected than *AP1* expression (normalized *AP1* expression change in the *soc1* mutant compared to wild type is two times that of *LFY*; SI Fig. 3-4). This illustrates that the effect of perturbations can considerably grow in magnitude throughout the network.

Regulation of AP1 by LFY

As mentioned above, for the regulation of *AP1* by LFY our initial analysis using the PCR time-course data indicated that we needed to introduce DNA-binding cooperativity in the equations in order to get a reasonable fit of the data. There is indeed experimental evidence for cooperativity in the LFY – *AP1* interaction, based on the LFY protein-DNA structure and additional experimental data [47]. In our modelling approach, cooperativity is defined by a Hill coefficient $n>1$ in the term in the differential equation describing the regulation of *AP1* by LFY. For the regulation of *AP1* by LFY, setting the value of $n=3$ resulted in a markedly improved fit of the wild type time-course data (SI Fig. 5). No improvement of the fit could be obtained for the other interactions by the introduction of a Hill coefficient larger than 1, meaning that the data does not contain evidence for cooperativity in those interactions. Cooperativity in the LFY – *AP1* interaction provides an additional predicted mechanism by which a small change in LFY, can lead to a large change in *AP1* expression. Experimental evidence indeed suggests that cooperativity in LFY binding to the AP1 promoter is important [47].

Regulation of LFY by AGL24 and SOC1

It has been suggested that SOC1 requires dimerization with AGL24 for binding to the *LFY* promoter. This is based on several sources of experimental evidence: (I) in yeast-two-hybrid assay, AGL24 and SOC1 form a heterodimer [48]; (II) SOC1 is only detected in the nucleus when AGL24 is present as well [39]; (III) *LFY* is expressed only in those tissues where *SOC1* and *AGL24* expression overlap [39]. Nevertheless, there is a significant difference between the flowering time observed for *soc1* and *agl24* mutants (Figure 4A). If these two proteins would bind the *LFY* promoter as AGL24-SOC1 dimer only, then knockout mutations in either *AGL24* or *SOC1* would equally reduce the dimer concentration; therefore, one would expect the same effect on *LFY*.

Based on these considerations, in our final model, AGL24 and SOC1 have independent roles in regulating *LFY*. We tested an alternative model version in which AGL24 and SOC1 only regulated *LFY* as a dimer and not separately from each other. This resulted in a decreased goodness-of-fit in particular for *LFY* (nrmse 43% instead of 7%) and in this alternative model, indeed the effect of *agl24* and *soc1* mutation on LFY and on flowering time were comparable, which contradicts available experimental data.

In our model, in which AGL24 and SOC1 have independent roles in regulating *LFY*, the simulated *LFY* expression is reduced by only ~25% in the *agl24* knockout mutant relative to its time-course expression in wild type. In contrast, *LFY* expression is reduced by ~65% in the *soc1* mutant (Figure 4B; SI Fig. 3). These predicted changes are consistent with what is experimentally observed in the microarray data for *LFY* (Figure 4C). If the expression level of *SOC1* in wild type would be much higher than that of *AGL24*, a hypothesis could be that elimination of such more abundant factor would

have a larger effect. However, in our expression data, expression levels of the two genes are of the same order of magnitude. According to the model, two parameters are important in describing the regulatory effect of SOC1 and AGL24 on *LFY*: DNA binding efficiency (represented by parameter K) and expression activation strength (parameter β). A difference in any of these two parameters between SOC1 and AGL24 could lead to a difference in the effect of *SOC1* versus *AGL24* mutation. In the set of parameter values we obtained for our model, the DNA binding efficiency for AGL24 (K_{10}) and SOC1 (K_{11}) binding to the *LFY* promoter is quite similar. However, there is a substantial difference in activation strength (β_7 vs β_6), with SOC1 being much more able to activate *LFY*, resulting in a much larger effect of *soc1* mutation compared to *agl24* mutation. Analysis of predicted flowering times for a range of values of β for SOC1 and AGL24 confirms the dependency on the SOC1 activation strength (Figure 4D). In addition, the flowering time observed for the double mutant *soc1/agl24* suggests a small additive effect when both genes are simultaneously knocked-out (Figure 4A). In agreement with that observation, the model simulation predicts a small additional reduction in *LFY* expression for the *soc1/agl24* double mutant (~80% vs. ~65% in single mutant; Figure 4B). Overall, these examples demonstrate how we get quantitative insight in the spread of perturbations through the network. Moreover, this demonstrates that we can analyse how the quantitative output of the network as a whole is governed by specific molecular interactions that build up the network.

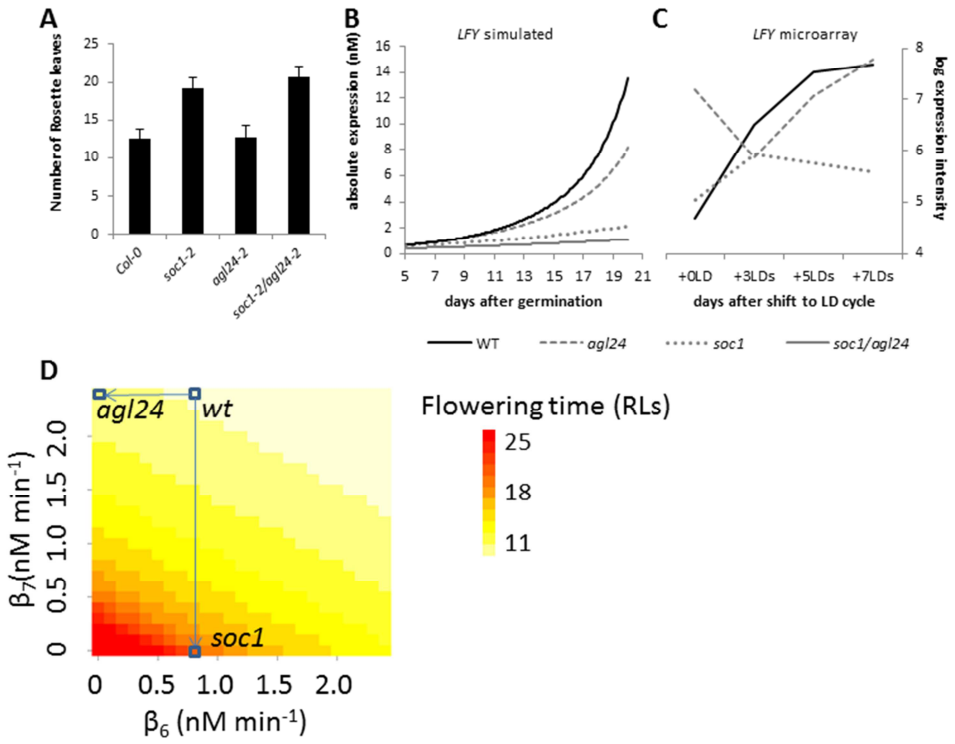


Figure 4: Effect of knockout mutations (*agl24*, *soc1* and *soc1/agl24*) on *LFY* expression and on flowering time. (A) Number of rosette leaves counted at the onset of flowering for wild type and mutants. The plants were grown in long-day conditions at 23°C. **(B-C)** *LFY* expression in wild type and mutants from simulations (B) or microarray experiments (C). The simulations show the expression time-course over 20 days after germination;

the microarray data consist of four time-points after transfer of plants grown in short-day to long-day conditions. **(D)** Effect of efficiency by which *LFY* expression is activated by AGL24 (β_6) and SOC1 (β_7), on predicted flowering time. Flowering time, predicted flowering time for given values of parameters. Blue boxes in heatmap indicate best-fit model parameters and the two mutants *soc1* and *agl24*; arrows point from best-fit model to mutants.

Discussion

An important reason to apply computational models to a biological system, such as the floral integration network, is that it allows investigating how the various interactions that together constitute the network, transmit perturbations into a final readout. Indeed, by integrating experimental data with modeling we analyse how different components of the flowering time regulation network react to changes in other components, finally leading to a specific flowering time. We specifically analysed the regulation of *LFY* by SOC1, the regulation of *LFY* by AGL24, and the regulation of *AP1* by LFY. In these cases, the activation strength was found to be the most important cause of the observed differences in magnitude of effect of perturbations, according to the model. This could mean that the protein with the higher predicted activation strength itself is a stronger transcriptional activator than the other protein, or it could indicate involvement in a protein interaction with a partner (not explicitly included in the model), which is a stronger activator. In the case of the different effect of the *soc1* mutation compared to the *agl24* mutation, it is important to consider that both SOC1 and AGL24 are known to form additional complexes, and such dimers might also play a role in their differential functioning [48]. In addition, as a general note on our interpretation of parameter values, it is important to realize that we use a fixed conversion of mRNA levels to protein levels; this means that potential differences in e.g. translation rate could complicate the interpretation of the parameters.

Previous work on modelling flowering time used the concepts of "photothermal units" or variants thereof as a way to computationally investigate flowering time and how it is influenced by the environment; as recently demonstrated, such models can in principle be connected to genetic information [49]. However, this does not provide a direct way to incorporate the regulatory interactions between genes, which are key towards a mechanistic understanding of flowering time regulation. Our work is more comparable to recent approaches, which start with defining interactions in a gene regulatory network and then develop a model based on this network [19,20]. Our approach extends the recently published Arabidopsis flowering time model [20] by fitting model parameters using dynamic expression data. The model by Jaeger et al. predicted that the AP1 time course does not show a sharp transition from low expression to the on-state, as seen in our experimental data. This indicates that a model, in which parameters are estimated purely based on mutant flowering times, might miss important aspects of gene expression dynamics. Additional time course data could in the future be obtained at various experimental conditions (temperature, light) as a step towards including

the effect of such conditions on the model. A direct advantage is that our parameters have a physical interpretation (e.g. activation strength, cooperativity, etc).

When analysing for which genes the model predictions were of better quality, the effects of an *SVP* overexpression mutant and an *FT* knock-out mutant on flowering time were predicted less accurately compared to other mutants (including *SVP* knock-out and *FT* overexpression mutants). *FT* and *SVP* are connected to each other in the network, which could indicate that in this part of the network the model needs refinement. In particular, given that *SVP* overexpression results in lower *FT* expression, the fact that both *SVP* overexpression and *FT* knock-out were not well predicted indicates that the effect of lower *FT* levels, either directly on *AP1* or more indirect via *SOC1*, is not perfectly captured. It is however also important to consider that the *FT* levels used as input in our model are relatively low, which is related to the fact that they are not measured at the peak of diurnal expression of *FT*. Another aspect to consider is that *FLC* and *SVP* are present as external inputs in the model and are not directly modelled; if a mutation in one of these impacts the other as well, the model would miss such effect, which would deteriorate prediction performance. This might indeed be the case, according to ChIP-seq data [35,36].

Clearly, there are several directions to expand our work. We do not specifically represent protein and RNA separately; currently the state of the art in the proteomics field does not allow high-throughput and precise quantification of protein levels during the vegetative phase of plant development. Recent evidence indicates however that for at least one component in the model, SVP, the effect of protein stability is important [50]. In theory, for the differential effects of *soc1* vs. *agl24* mutation, for which we provide an explanation in terms of a difference in a specific parameter in the model, difference in protein levels in spite of similarity in RNA levels could also be relevant, although there is currently no experimental data that indicates this.

In general, the amount of detail in the model will always be a compromise. This holds as well for the type of interactions in the network. Currently, regulatory interactions are modelled, whereas protein-protein interactions are not explicitly included. Nevertheless, the way in which regulatory inputs are combined gives an implicit representation of the way in which proteins interact with each other. Although the importance of complex formation for the components of the network is clear [48,51], one reason why at our level of detail they can be excluded might be that they are mainly relevant for specifying the correct regulatory interactions (which are explicitly defined in the model equations) and less so for the kinetics of the model. Depending on the availability of proteomics data, it would however be straightforward to include e.g. protein dimerization explicitly in our equations.

Currently, we focused on a core set of genes involved in integrating various flowering time signals. Given that input from the environment converges on various components of the flowering integration

network, an exciting follow-up step will be to incorporate environmental cues as the next layer of information in the gene regulatory network. This could include both direct environmental effects on some of the model components, or modelling complete upstream pathways. As an example of direct environmental influence that could be modelled, recent data indicates that the above mentioned effects of SVP protein stability as well as FLM alternative splicing depend on temperature [52]. To include the former, although protein levels are not explicitly present in our model, an effect of temperature on stability could be represented by changing the *SVP* decay parameter; for FLM, additional equations describing the two isoforms would be needed. As for the modelling of upstream pathways, in a recent overview of known effects of mutations, ~150 genes were listed as being currently known to impact flowering time [53]. It remains to be seen which would be the best approach to include such genes and whether it is essential to include all of them for reliable predictions. Given sufficient time-course data it might be possible to use the same approach as presented here. However, it would also be an option to focus detailed modelling efforts on particular parts of the network; for example, for the influence of light on the circadian clock, models have already been developed [4-6] and these could be connected to our model. Other parts of the network could be treated in a more coarse grained, statistical approach.

To conclude, we present a dynamic and predictive model for flowering time regulation. Our work presents a framework for studying the mechanisms of flowering time regulation, by addressing how different quantitative inputs are combined into a single quantitative output, the timing of flowering.

Methods

Plant materials and growth conditions

For the time-course gene expression studies Arabidopsis Col-0 wild type plants were grown under long-day conditions (16 hrs light, 8 hrs dark; 21 °C) on rockwool and received 1 g/L Hyponex^{TM} plant food solution two times per week. Rosette leaves and shoot apical meristem enriched material was harvested daily at ZT3 from seven plants per sample in duplicate.

Plants for flowering time analysis were grown in growth chambers with controlled environment (23ºC, 65% relative humidity) under long-day conditions (16 hrs light, 8 hrs dark). Plants were raised on soil under a mixture of Cool White and Gro-Lux Wide Spectrum fluorescent lights, with a fluorescence rate of 125 to 175 mmol m^{-2} s^{1}. For flowering time measurements, the total number of primary rosette leaves was scored at visual bolting. The position of the plants from the different genotypes were randomized in the trays, and the flowering time phenotype was recorded without prior knowledge of the genotype. Plants for microarray experiments were grown on soil in growth chambers (23 °C, 65% relative humidity) under short-day conditions (8 hrs light, 16 hrs dark) for 25 days (Col-0, *soc1-6* (SALK_138131), *agl24* (SALK_095007), *flc-3*) or 28 days (*fd-3*). Flowering was induced by shifting

plants to long-day conditions (16 hrs light, 8 hrs dark).

Quantitative qRT-PCR data

RNA was isolated from the plant samples (max 100 mg of grinded plant material) using the InviTrap Spin Plant RNA Mini Kit. Subsequently, a DNAse (Invitrogen) treatment was performed, which was stopped with 1µL of a 20 mM EDTA solution and 10 minutes incubation at 65°C. Total RNA concentration was measured, and 1 µg RNA was used to perform cDNA synthesis by the Taqman MultiScribe™ Reverse Transcriptase kit (LifeTechnologies). qRT-PCR was performed with the SYBR green mix from BioRad using the gene specific oligonucleotides indicated in SI table 4. *YELLOW-LEAF-SPECIFIC GENE8 (YLS8)* was implemented as reference gene for the analyses.

The relative gene expression was given by $E_{target} = 2^{\Delta Ct}$, where Ct stands for the threshold cycle and $\Delta Ct = Ct_{target} - Ct_{reference}$. From that, the absolute abundance was estimated by $A_{target} = E_{target} \times s$, where s stands for a scaling factor obtained by dividing the average abundance that a transcript reaches in a cell by the highest E_{target}value among all samples, and multiplying by an assumed maximal protein abundance. Since a linear relationship between abundances of RNA and protein is assumed in the model, the average transcript abundance was adjusted based on average abundance of a protein in cell. An available estimate for the range of protein abundance is between 400nM and 1400nM [54]. From this range, the average abundance for the flowering time gene products was arbitrarily chosen (500nM). This means that the maximum absolute expression among all samples is equal to 500nM (Fig. 2).

Microarray data

Microarray time series experiments were performed as previously described [55] using RNA isolated from manually dissected shoot apices of Col-0, *soc1-6*, *agl24*, and *fd-3*. Briefly, biotinylated probes were prepared from 1 µg of total RNA using the MessageAmp II-Biotin Enhanced Kit (Ambion) following the manufacturer's instructions and hybridized to Arabidopsis ATH1-121501 gene expression array (Affymetrix). Arrays were washed on a GeneChip Fluidics Station 450 (affymetrix) and scanned on an Affymetrix GeneChip Scanner 7G using default settings. Expression data for Col-0, *soc-6*, *agl24*, and *fd-3* have been deposited with ArrayExpress (E-MEXP-4001). Expression data for *flc-3* (ArrayExpress: E-MEXP-2041) have previously been published [56]. The probe intensities were normalized and the gene expression estimates were obtained using the gcRMA library of R/Bioconductor [57].

The model

The regulatory interactions shown in Figure 1 were modelled by equations based on Hill kinetics. It was assumed that spatial aspects could be ignored (except for FT transport); hence the interactions

between the components are described by a set of ordinary differential equations (ODEs). Furthermore, only proteins were explicitly modelled, and a linear relationship between RNA levels and protein levels was assumed. The model is composed of the following equations:

(1) $\frac{dx_{FT}}{dt} = \beta_1 \left(\frac{K_1}{K_1+x_{SVP,l}}\right)\left(\frac{K_2}{K_2+x_{FLC,l}}\right) - d_1 x_{FT}$

(2) $\frac{dx_{AGL24}}{dt} = \beta_2 \frac{x_{SOC1}}{K_3+x_{SOC1}} - d_2 x_{AGL24}$

(3) $\frac{dx_{SOC1}}{dt} = \left[\left(\frac{\beta_3 x_{AGL24}}{K_4+x_{AGL24}}\right) + \left(\frac{\beta_4 x_{SOC1}}{K_5+x_{SOC1}}\right) + \left(\frac{\beta_5 x_{FT,t-\Delta}}{K_6+x_{FT,t-\Delta}}\right)\left(\frac{x_{FD}}{K_7+x_{FD}}\right)\right]\left(\frac{K_8}{K_8+x_{SVP,m}}\right)\left(\frac{K_9}{K_9+x_{FLC,m}}\right) - d_3 x_{SOC1}$

(4) $\frac{dx_{LFY}}{dt} = \left[\left(\frac{\beta_6 x_{AGL24}}{K_{10}+x_{AGL24}}\right) + \left(\frac{\beta_7 x_{SOC1}}{K_{11}+x_{SOC1}}\right) + \left(\frac{\beta_8 x_{AP1}}{K_{12}+x_{AP1}}\right)\right] - d_4 x_{LFY}$

(5) $\frac{dx_{AP1}}{dt} = \left(\frac{\beta_9 x_{LFY}^n}{K_{13}^n+x_{LFY}^n}\right) + \left(\frac{\beta_{10} x_{FT,t-\Delta}}{K_{14}+x_{FT,t-\Delta}}\right) + \left(\frac{\beta_{11} x_{FD}}{K_{15}+x_{FD}}\right) - d_5 x_{AP1}$

(6) $\frac{dx_{FD}}{dt} = \left(\frac{\beta_{12} x_{LFY}}{K_{16}+x_{LFY}}\right) - d_6 x_{FD}$

For *FLC* and *SVP*, gene expression is represented in the leaves ($x_{FLC,l}$ and $x_{SVP,l}$) and meristem ($x_{FLC,m}$ and $x_{SVP,m}$). For all the other genes, the variables correspond to expression in the meristem. Note that for *SVP* and *FLC* there are no equations; they act as external inputs in the model, and their regulation is not explicitly modelled. The parameters in the equations have the following meaning (see SI Table 2-3 for further details): parameters β and K stand for the maximum transcription rate and for the abundance at half-maximum transcription rate, respectively; d_i stands for the degradation rate of the products of gene i; Δ stands for the time needed for transporting FT from the leaves to the meristem; $x_{FT,t-\Delta}$ is the amount of FT in the meristem at time t which is assumed to be equal to that in the leaves at time t-Δ; and n is the Hill coefficient describing cooperativity in the regulation of *AP1* by LFY.

Equations (1-5) are based on the following specific assumptions: (I) *SVP* and *FLC* bind to *FT* and *SOC1* promoters as a dimer. This is implicitly represented by the multiplication of the Hill terms associated to the S*VP*- and *FLC*-mediated regulations of *FT* and *SOC1*. (II) FD requires dimerization

with FT in order to activate *SOC1* expression. (III) FD can activate *AP1* as a monomer. (IV) Recently it was shown that in rice the interaction between FT and FD is bridged by a 14-3-3 protein [58] and probably this is also the case in Arabidopsis; nevertheless, we did not include 14-3-3s in our model, because these proteins seem to be highly abundant and hence not limiting for floral induction. The specific form of the equations and the assumptions that they represent were adjusted by assessing the fitting and the flowering time predictions of variants for the five equations. In addition, to obtain a good fit for the equation associated to *AP1*, the degree of cooperativity (*n*) for the LFY-mediated regulation of *AP1* was set to *n*=3.

Parameter estimation

In the model, the expression dynamics x_i of a gene i depends on the parameter values associated to $\frac{dx_i}{dt}$ and on the expression values over the time-course of the direct regulators of i. To independently fit an equation $\frac{dx_i}{dt}$ to its corresponding time-course, the expressions of the direct regulators of i in the right-hand side of the equation were taken from the data, and interpolated with a polynomial fit. This decoupling method has previously been described in full detail [14]. By applying this method, it is possible to find the parameters for each equation without knowing the parameters associated to the other equations; thus, alleviating the high computational demand put on the search algorithm by the total number of parameters. This optimization step was carried out by the *MultiStart* solver implemented in MATLAB (R2012a, The MathWorks UK, Cambridge). The parameters were then input in the whole systems of equations as starting point for a second optimization step. In this second step, the equations were solved as a system and the expressions of the direct regulators of i were taken from their associated ordinary differential equation solutions. This was carried out by the *lsqnonlin* solver (implemented in MATLAB) to fine-tune the fitting obtained by the first optimization step.

To assess the goodness of fit for each gene, the normalized root mean square error (NRMSE) was used, which equals $\frac{RMSE}{x_{max}-x_{min}}$, with RMSE equal to $\sum_{i=1}^{T}\sqrt{\frac{\sum (x_i^{exp}-x_i^{pred})^2}{n}}$; here x_{max}, x_{min} are the maximum and minimum observed expression value; x_i^{exp} and x_i^{pred} are the experimental and predicted values at time *i*; *T* is the total number of timepoints, and the sum is over all timepoints.

Model simulations

The equations were solved using MATLAB, integrated with the stiff solver *ode23s*. For simulations of gene expressions of Arabidopsis wild type grown at 23°C/LD, the initial gene abundances were taken equal to the first expression time-points and the parameters were the same as described in SI Table 3. To simulate gene expression in mutants, the expression associated to a mutated gene i was fixed to a constant value $x_i = k_{mut}$. For the knock-out null mutants (*ft-10*, *fd-3*, *flc-3*), the values of k_{mut} were adjusted to zero; and for the knockdown mutants (not null mutations), k_{mut} values were adjusted to a

small percentage of the expression of i observed in the first time-point from wild type Col-0 (SI Table 5). For the overexpression mutants, the values of k_{mut} were set to five times the maximum absolute expression among all samples (2500nM).

To assess the model predictions of changes in gene expression we compared predicted relative changes with relative changes obtained with microarray data. To do so, we calculated the predicted total amount of expression (integral of the predicted time-course from day 0 to day 20) using the *trapz* function in MATLAB. Subsequently, these values were scaled by substracting the wild type value and then dividing by the wild type value. Similarly, the experimental relative change was calculated based on the microarray data. Note that comparing those values focusses on the effect of a mutation on dynamics of genes in the network over the complete time-course and as such takes into account the fact that the experimental conditions of the microarray experiment cannot directly be simulated (flowering-inducing shift from short-day to long-day conditions).

Model predictions of flowering time

The predictions of flowering time were based on *AP1* expression. For that it was assumed that, at a molecular level, *Arabidopsis* undergoes the floral transition in the moment that *AP1* expression initiates. Therefore, according to our experimental *AP1* time series, for wild type Col-0, the floral transition takes place between days 12 and 13 after germination. For simplification, we take the exact day 12.6 because it corresponds to the average number of rosette leaves (RLs) observed at the onset of flowering for wild type Col-0. To estimate the flowering time from mutant simulations, we use the time in which *AP1* expression reaches the same simulated expression value as obtained at day 12.6 for wild type Col-0. This implies that the *AP1* expression threshold for triggering the floral transition is the same for different plant growth conditions and mutants. Because flowering times are usually reported in number of rosette leaves (RLs) we subsequently scaled the predicted days to RLs by assuming a linear relationship between the number of RLs observed at the onset of flowering and the time in days after germination that *Arabidopsis thaliana* undergoes the floral transition at a molecular level.

In addition to the set of mutants obtained in consistent conditions in this work, we also included existing mutant data. Wild type Col-0 flowering time in these experiments is somewhat different from that observed in our experiments. In addition, flowering times in literature are mostly reported in rosette leaves (RL), and not directly in days. To be able to integrate those data, we scaled existing mutant data with a linear factor which is chosen in such a way as to scale the wild type Col-0 flowering time to 12.6 RL.

Author summary

A major goal of computational biology is to understand how gene regulatory networks, in which genes regulate each other's expression, result in a particular output. Here, we address this problem in the context of flowering time, a very important trait of plants: flowering at the right time is essential for reproduction. We study a core integration network consisting of eight transcription factors involved in flowering time regulation in the model plant *Arabidopsis thaliana*. We obtained various experimental datasets in order to estimate model parameters and in order to validate the model. This included gene expression time courses in wild type plants, flowering times of wild type and mutants, and expression data in mutant backgrounds. Our model enables to predict and understand how changes in one or more of the genes in the network influence flowering time. The ability to predict flowering time changes may benefit future breeding efforts aimed at for example mitigating the effects of climate change.

Funding statement

This work was supported by a Max Planck postdoctoral fellowship to DP, an EMBO Long-Term postdoctoral fellowship to MCK, a Deutsche Forschungsgemeinschaft (DFG) grant (ERA-PG "BloomNet", SCHM1560/7-1) to MS and the Max Planck Society, a Netherlands Organisation for Scientific Research (NWO) VENI grant (863.08.027) to ADJvD, the SYSFLO Marie Curie Initial Training Network (FLV), and by the Netherlands Consortium for Systems Biology. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgement

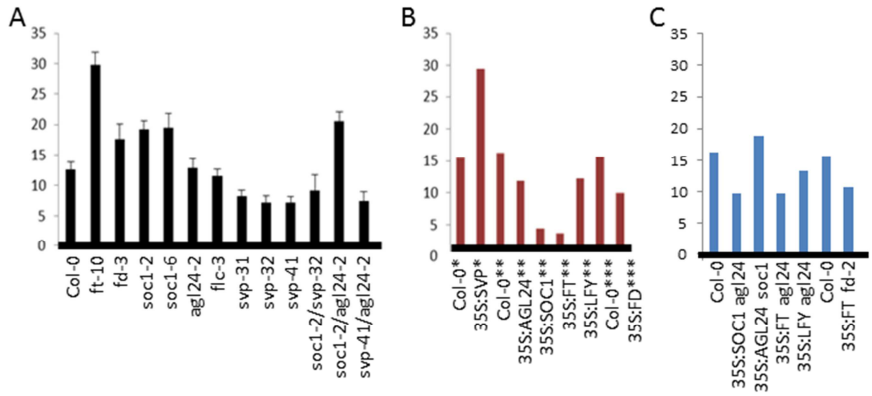
We thank Levi Yant for help with *fd-3* microarray experiments.

References

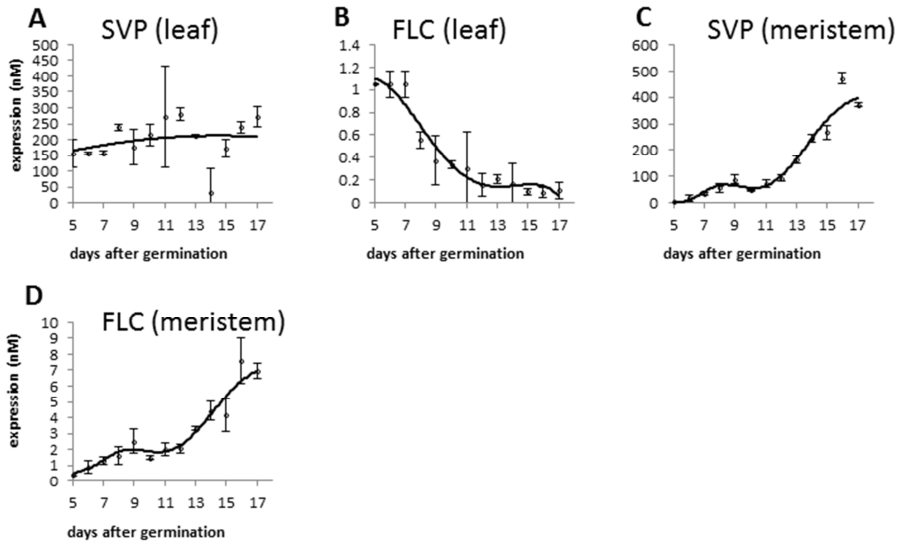
1. Srikanth A, Schmid M (2011) Regulation of flowering time: all roads lead to Rome. Cell Mol Life Sci 68: 2013-2037.
2. Andres F, Coupland G (2012) The genetic basis of flowering responses to seasonal cues. Nat Rev Genet 13: 627-639.
3. Chew YH, Smith RW, Jones HJ, Seaton DD, Grima R, et al. (2014) Mathematical models light up plant signaling. Plant Cell 26: 5-20.
4. Gould PD, Ugarte N, Domijan M, Costa M, Foreman J, et al. (2013) Network balance via CRY signalling controls the Arabidopsis circadian clock over ambient temperatures. Mol Syst Biol 9: 650.
5. Keily J, Macgregor DR, Smith RW, Millar AJ, Halliday KJ, et al. (2013) Model selection reveals control of cold signalling by evening-phased components of the plant circadian clock. Plant J.
6. Pokhilko A, Fernandez AP, Edwards KD, Southern MM, Halliday KJ, et al. (2012) The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops. Mol Syst Biol 8: 574.
7. Schmal C, Reimann P, Staiger D (2013) A Circadian Clock-Regulated Toggle Switch Explains AtGRP7 and AtGRP8 Oscillations in Arabidopsis thaliana. PLoS Comput Biol 9.
8. Jonsson H, Heisler MG, Shapiro BE, Meyerowitz EM, Mjolsness E (2006) An auxin-driven polarized transport model for phyllotaxis. Proc Natl Acad Sci U S A 103: 1633-1638.
9. van Mourik S, Kaufmann K, van Dijk AD, Angenent GC, Merks RM, et al. (2012) Simulation of organ patterning on the floral meristem using a polar auxin transport model. PLoS One 7: e28762.
10. Peret B, Middleton AM, French AP, Larrieu A, Bishopp A, et al. (2013) Sequential induction of auxin efflux and influx carriers regulates lateral root emergence. Mol Syst Biol 9: 699.
11. Grieneisen VA, Xu J, Maree AF, Hogeweg P, Scheres B (2007) Auxin transport is sufficient to generate a maximum and gradient guiding root growth. Nature 449: 1008-1013.
12. Salazar JD, Saithong T, Brown PE, Foreman J, Locke JC, et al. (2009) Prediction of photoperiodic regulators from quantitative gene circuit models. Cell 139: 1170-1179.
13. Song YH, Smith RW, To BJ, Millar AJ, Imaizumi T (2012) FKF1 conveys timing information for CONSTANS stabilization in photoperiodic flowering. Science 336: 1045-1049.
14. van Mourik S, van Dijk AD, de Gee M, Immink RG, Kaufmann K, et al. (2010) Continuous-time modeling of cell fate determination in Arabidopsis flowers. BMC Syst Biol 4: 101.
15. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER (2004) A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. Plant Cell 16: 2923-2939.
16. Mendoza L, Alvarez-Buylla ER (1998) Dynamics of the genetic regulatory network for Arabidopsis thaliana flower morphogenesis. J Theor Biol 193: 307-319.
17. Sanchez-Corrales YE, Alvarez-Buylla ER, Mendoza L (2010) The Arabidopsis thaliana flower organ specification gene regulatory network determines a robust differentiation process. J Theor Biol 264: 971-983.
18. Jung C, Muller AE (2009) Flowering time control and applications in plant breeding. Trends Plant Sci 14: 563-573.
19. Dong Z, Danilevskaya O, Abadie T, Messina C, Coles N, et al. (2012) A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. PLoS One 7: e43450.
20. Jaeger KE, Pullen N, Lamzin S, Morris RJ, Wigge PA (2013) Interlocking feedback loops govern the dynamic behavior of the floral transition in Arabidopsis. Plant Cell 25: 820-833.
21. Satake A, Kawagoe T, Saburi Y, Chiba Y, Sakurai G, et al. (2013) Forecasting flowering phenology under climate warming by modelling the regulatory dynamics of flowering-time genes. Nat Commun 4: 2303.
22. Welch SM, Roe JL, Dong ZS (2003) A genetic neural network model of flowering time control in Arabidopsis thaliana. Agronomy Journal 95: 71-81.
23. Pose D, Yant L, Schmid M (2012) The end of innocence: flowering networks explode in complexity. Curr Opin Plant Biol 15: 45-50.
24. Jang S, Torti S, Coupland G (2009) Genetic and spatial interactions between FT, TSF and SVP during the early stages of floral induction in Arabidopsis. Plant J 60: 614-625.
25. Helliwell CA, Wood CC, Robertson M, James Peacock W, Dennis ES (2006) The Arabidopsis FLC protein interacts directly in vivo with SOC1 and FT chromatin and is part of a high-molecular-weight protein complex. Plant J 46: 183-192.
26. Li D, Liu C, Shen L, Wu Y, Chen H, et al. (2008) A repressor complex governs the integration of flowering signals in Arabidopsis. Dev Cell 15: 110-120.
27. Tao Z, Shen L, Liu C, Liu L, Yan Y, et al. (2012) Genome-wide identification of SOC1 and SVP targets during the floral transition in Arabidopsis. Plant J 70: 549-561.
28. Mathieu J, Warthmann N, Kuttner F, Schmid M (2007) Export of FT protein from phloem companion cells is sufficient for floral induction in Arabidopsis. Curr Biol 17: 1055-1060.
29. Corbesier L, Vincent C, Jang S, Fornara F, Fan Q, et al. (2007) FT protein movement contributes to long-distance signaling in floral induction of Arabidopsis. Science 316: 1030-1033.
30. Abe M, Kobayashi Y, Yamamoto S, Daimon Y, Yamaguchi A, et al. (2005) FD, a bZIP protein mediating signals from the floral pathway integrator FT at the shoot apex. Science 309: 1052-1056.
31. Wigge PA (2013) Ambient temperature signalling in plants. Curr Opin Plant Biol.
32. Yoo SK, Chung KS, Kim J, Lee JH, Hong SM, et al. (2005) CONSTANS activates SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 through FLOWERING LOCUS T to promote flowering in Arabidopsis. Plant Physiol 139: 770-778.

33. Wigge PA, Kim MC, Jaeger KE, Busch W, Schmid M, et al. (2005) Integration of spatial and temporal information during floral induction in Arabidopsis. Science 309: 1056-1059.
34. Hartmann U, Hohmann S, Nettesheim K, Wisman E, Saedler H, et al. (2000) Molecular cloning of SVP: a negative regulator of the floral transition in Arabidopsis. Plant J 21: 351-360.
35. Gregis V, Andres F, Sessa A, Guerra RF, Simonini S, et al. (2013) Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in Arabidopsis. Genome Biol 14: R56.
36. Deng W, Ying H, Helliwell CA, Taylor JM, Peacock WJ, et al. (2011) FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of Arabidopsis. Proc Natl Acad Sci U S A 108: 6680-6685.
37. Immink RG, Pose D, Ferrario S, Ott F, Kaufmann K, et al. (2012) Characterization of SOC1's central role in flowering by the identification of its upstream and downstream regulators. Plant Physiol 160: 433-449.
38. Michaels SD, Ditta G, Gustafson-Brown C, Pelaz S, Yanofsky M, et al. (2003) AGL24 acts as a promoter of flowering in Arabidopsis and is positively regulated by vernalization. Plant J 33: 867-874.
39. Lee J, Oh M, Park H, Lee I (2008) SOC1 translocated to the nucleus by interaction with AGL24 directly regulates leafy. Plant J 55: 832-843.
40. Wagner D, Sablowski RW, Meyerowitz EM (1999) Transcriptional activation of APETALA1 by LEAFY. Science 285: 582-584.
41. Kaufmann K, Wellmer F, Muino JM, Ferrier T, Wuest SE, et al. (2010) Orchestration of floral initiation by APETALA1. Science 328: 85-89.
42. Hempel FD, Weigel D, Mandel MA, Ditta G, Zambryski PC, et al. (1997) Floral determination and expression of floral regulatory genes in Arabidopsis. Development 124: 3845-3853.
43. Giakountis A, Coupland G (2008) Phloem transport of flowering signals. Curr Opin Plant Biol 11: 687-694.
44. Torti S, Fornara F, Vincent C, Andres F, Nordstrom K, et al. (2012) Analysis of the Arabidopsis shoot meristem transcriptome during floral transition identifies distinct regulatory patterns and a leucine-rich repeat protein that promotes flowering. Plant Cell 24: 444-462.
45. Michaels SD, Amasino RM (1999) FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell 11: 949-956.
46. van Leeuwen W, Ruttink T, Borst-Vrenssen AW, van der Plas LH, van der Krol AR (2001) Characterization of position-induced spatial and temporal regulation of transgene promoter activity in plants. J Exp Bot 52: 949-959.
47. Hames C, Ptchelkine D, Grimm C, Thevenon E, Moyroud E, et al. (2008) Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. EMBO J 27: 2628-2637.
48. de Folter S, Immink RG, Kieffer M, Parenicova L, Henz SR, et al. (2005) Comprehensive interaction map of the Arabidopsis MADS Box transcription factors. Plant Cell 17: 1424-1433.
49. Wilczek AM, Roe JL, Knapp MC, Cooper MD, Lopez-Gallego C, et al. (2009) Effects of genetic perturbation on seasonal life history plasticity. Science 323: 930-934.
50. Lee JH, Ryu HS, Chung KS, Pose D, Kim S, et al. (2013) Regulation of temperature-responsive flowering by MADS-box transcription factor repressors. Science 342: 628-632.
51. Immink RG, Tonaco IA, de Folter S, Shchennikova A, van Dijk AD, et al. (2009) SEPALLATA3: the 'glue' for MADS box transcription factor complex formation. Genome Biol 10: R24.
52. Pose D, Verhage L, Ott F, Yant L, Mathieu J, et al. (2013) Temperature-dependent regulation of flowering by antagonistic FLM variants. Nature.
53. Lloyd J, Meinke D (2012) A comprehensive dataset of genes with a loss-of-function mutant phenotype in Arabidopsis. Plant Physiol 158: 1115-1129.
54. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, et al. (2003) Global analysis of protein expression in yeast. Nature 425: 737-741.
55. Schmid M, Uhlenhaut NH, Godard F, Demar M, Bressan R, et al. (2003) Dissection of floral induction pathways using global expression analysis. Development 130: 6001-6012.
56. Mathieu J, Yant LJ, Murdter F, Kuttner F, Schmid M (2009) Repression of flowering by the miR172 target SMZ. PLoS Biol 7: e1000148.
57. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
58. Taoka K, Ohki I, Tsuji H, Furuita K, Hayashi K, et al. (2011) 14-3-3 proteins act as intracellular receptors for rice Hd3a florigen. Nature 476: 332-335.

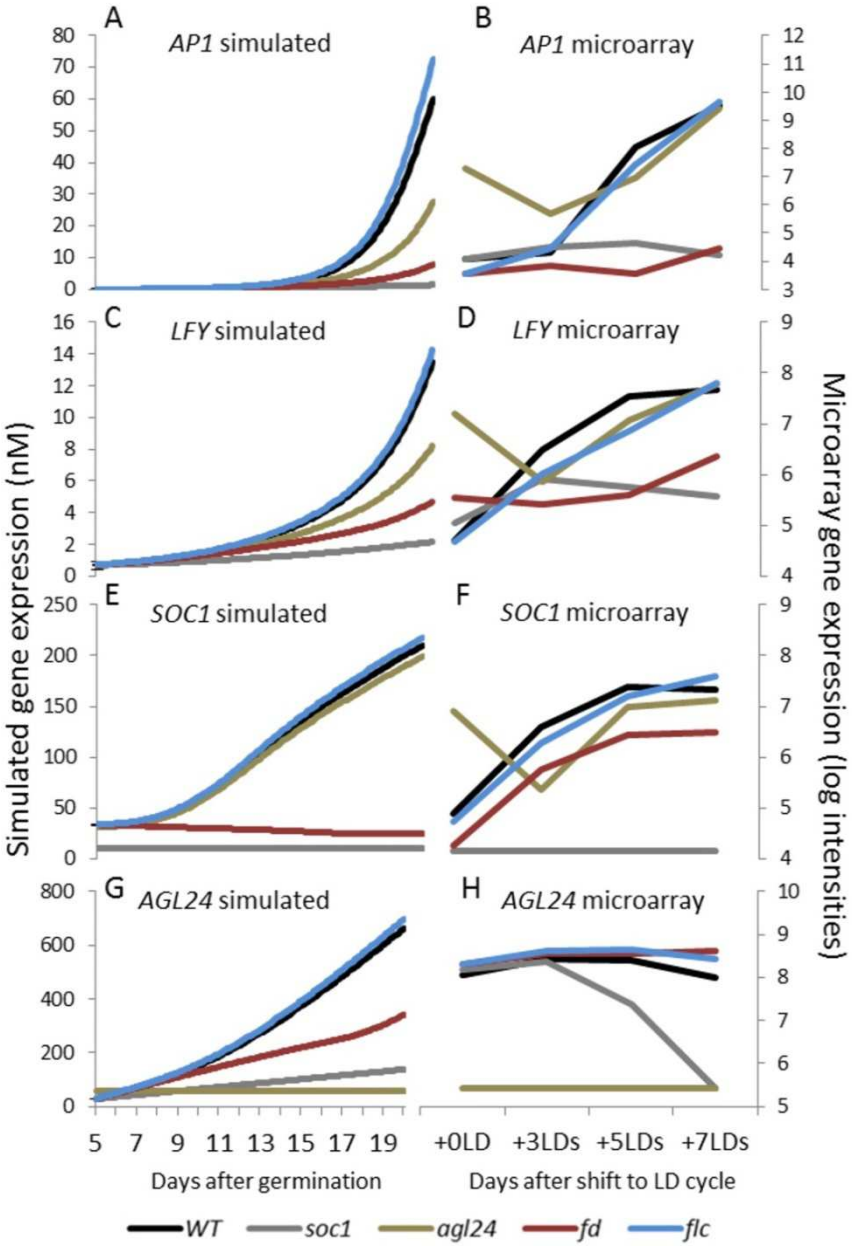
Supporting Information



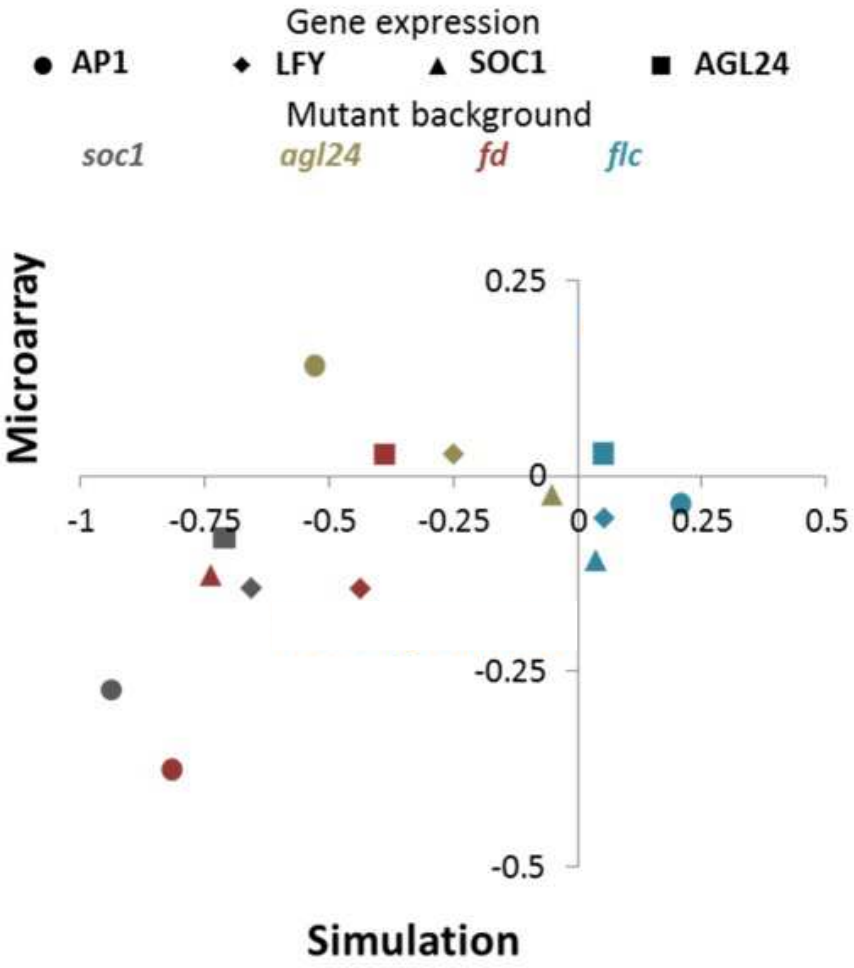
Supplemental Figure 1: Flowering time for Arabidopsis mutants. (A) Flowering time of knock-down/knock-out mutants measured in this work. Plants were grown in long-day conditions at 23°C. The standard deviation is indicated. (B) Flowering time for transgenic gene overexpression lines obtained from literature survey. Datasources: * **[59]**; ** **[60]**; and ***[33]. (C) Flowering time for mutants used in model validation**;** obtained from [59] except for 35S:FT fd-2 [33]. Although similar growth conditions are reported, the flowering time for wild type Col-0 varies among experiments. For this reason, when comparing model predictions with those data, we scaled the literature data based on the ratio between Col-0 flowering time observed in these experiments (panel B, C), and that observed in our experiment (panel A).



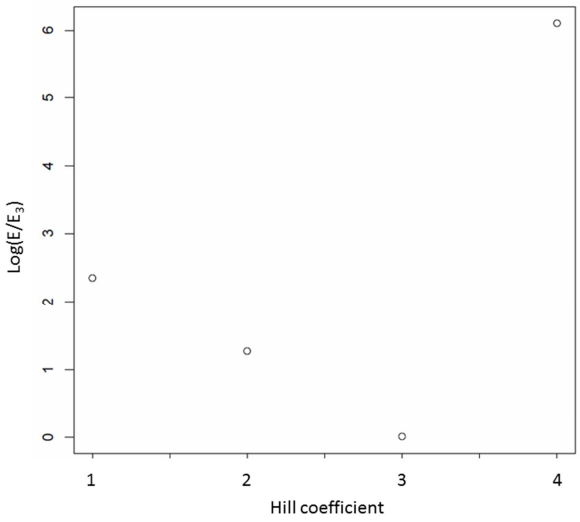
Supplemental Figure 2: Time-course expression of *FLC* and *SVP* in Arabidopsis wild-type (WT). Gene expression was measured by qRT-PCR (shown as dots) of wild type Col-0 plants grown under long-day conditions at 23°C (average and standard deviation are shown). Note that *FLC* and *SVP* are not regulated by other components of the network and hence are present as input factors, and their expression level is not simulated by the model. The continuous lines show interpolated gene expression which was used as the external input. qRT-PCR data indicated as meristem was obtained from meristem enriched material**.**



Supplemental Figure 3: Expression of *AP1*, *LFY*, *SOC1* and *AGL24* in *Arabidopsis* wild-type (WT) and four mutant backgrounds (*soc1*, *agl24*, *fd* and *flc*). Gene expression obtained either from simulations **(A,C,E,G)** or microarray experiments **(B,D,F,H)**. The simulations show the time-course over 20 days after germination, whilst the microarray data over four time-points after the flowering-inducing shift of plants grown in short-days conditions transferred to long-day conditions (time-points 0, 3, 5 and 7 days after the shift).



Supplemental Figure 4: Comparison of experimental and simulated changes in gene expressions for *AP1, LFY, SOC1* and *AGL24* in four mutant backgrounds (*soc1, agl24, fd* and *flc*). The difference in total gene expression between wild type (WT) and mutants was calculated using simulations (*x*-axis) and microarray data (*y*-axis). For each gene, the difference between its expression in WT and in mutant is given by the difference between the area under the WT time-course and that of mutants (SI Fig 3); then normalized against the area under the WT time-course. Each colour represents the comparison of one of the four mutants against WT; and each of the compared genes is represented by dots in different shapes. Positive values are obtained when an increase in expression is observed for a mutant compared to that in WT. Pearson's correlation coefficient between simulated and microarray values is 0.67.



Supplemental Figure 5: Fit of AP1 equation for various values of the Hill coefficient. Hill coefficient n was set to 1,2,3, or 4, and Error Function (E, sum of squared residuals) of resulting fit was divided by the Error Function obtained for n=3 (E_3). Value shown is $\log(E/E_3)$. Lower value indicates better fit.

Supplemental Table 1: Regulatory relationships among the flowering time transcription factors.[a]

TF	Activated by	Repressed by
SVP	NA	NA
FLC	NA	NA
AGL24	SOC1 [61]	NA
SOC1	FT [32], SOC1 [37], AGL24 [38] and FD [30]	SVP [26,34], FLC [36]
LFY	AGL24 [39], SOC1 [62] and AP1 [41]	NA
FT	NA	SVP [24] and FLC [25]
FD	LFY (Jaeger et al. 2013)	NA
AP1	LFY [40], FT and FD [30,33]	NA

a Note that only regulatory interactions that are relevant before or at the floral transition are included. For example, AP1 regulates additional components of the network after its expression has been initiated, but these interactions are not relevant for the timing of the initial up-regulation of AP1.

Supplemental Table 2: Description of the lower and upper bounds for the model parameters.

Parameter	Description	Principal determinants	Unit	Lower limit	Upper limit	Ref
β	Maximum transcription rate	transcriptional efficiency	$nM * min^{-1}$	0.001	200	a
K	Abundance at half-maximum transcription rate	binding interface of transcription factor/DNA	nM	0.001	2000	NA
d	Degradation rate of gene products	protein and RNA stability, pos-transcriptional/-translational regulation	$\min^{-1}$	0.001	1	b
Δ	Time needed for transporting FT from the leaves to the meristem		days	0	1	c

[a] Based on data in [63], we take [0.001, 200] $nM \times \min^{-1}$ as a reasonable range for the possible limit values of ß.

[b] A range for decay [10^{-3}, 10^{-1}] $\min^{-1}$ is given in [64]. We take [10^{-3}, 1] $\min^{-1}$ as a reasonable range for the possible limit values of decay.

[c] The range for the delay parameter is adjusted based on the assumption that FT reaches the meristem within at maximum 1 day after being translated in the leaves.

Supplemental Table 3: Model parameters estimated from experimental expression time-course data.

Parameters	Regulatory interaction / gene	Value	Unit
K_1	SVP → FT	0.63	nM
K_2	FLC → FT	985	nM
K_3	SOC1 → AGL24	125	nM
K_4	AGL24 → SOC1	1182	nM
K_5	SOC1 → SOC1	695	nM
K_6	FT → SOC1	4.8	nM
K_7	FD → SOC1	2.4	nM
K_8	SVP → SOC1	909	nM
K_9	FLC → SOC1	501	nM
K_{10}	AGL24 → LFY	1011	nM
K_{11}	SOC1 → LFY	842	nM
K_{12}	AP1 → LFY	346	nM
K_{13}	LFY → AP1	947	nM
K_{14}	FT → AP1	10.1	nM
K_{15}	FD → AP1	700	nM
K_{16}	LFY→ FD	7.9	nM
$\boldsymbol{\beta_1}$	SVP/FLC → FT	51	$nM * min^{-1}$
$\boldsymbol{\beta_2}$	SOC1→ AGL24	100	$nM * min^{-1}$
$\boldsymbol{\beta_3}$	AGL24 → SOC1	0.52	$nM * min^{-1}$
$\boldsymbol{\beta_4}$	SOC1 → SOC1	64	$nM * min^{-1}$
$\boldsymbol{\beta_5}$	FT/FD → SOC1	189	$nM * min^{-1}$
$\boldsymbol{\beta_6}$	AGL24 → LFY	0.79	$nM * min^{-1}$
$\boldsymbol{\beta_7}$	SOC1 → LFY	2.4	$nM * min^{-1}$
$\boldsymbol{\beta_8}$	AP1 → LFY	22	$nM * min^{-1}$
$\boldsymbol{\beta_9}$	LFY → AP1	99.8	$nM * min^{-1}$

$\boldsymbol{\beta_{10}}$	FT → AP1	10	$nM * min^{-1}$
$\boldsymbol{\beta_{11}}$	FD→ AP1	5.0	$nM * min^{-1}$
$\boldsymbol{\beta_{12}}$	LFY→ FD	8.5	$nM * min^{-1}$
d_1	FT	0.10	min^{-1}
d_2	AGL24	0.0010	min^{-1}
d_3	SOC1	0.11	min^{-1}
d_4	LFY	0.017	min^{-1}
d_5	AP1	0.86	min^{-1}
d_6	FD	0.0075	min^{-1}
Δ	FT	0.50	days
n	LFY→ AP1	3	-

Supplemental Table 4: Oligonucleotide sequences of oligonucleotides used in qRT-PCR experiments.

TF	AtG number	Sequence Forward oligonucleotide	Sequence Reverse oligonucleotide
SVP	At2G22540.1	PDS3106 5'- GAAGAGAACGAGCG ACTTGG-3'	PDS3107 5'- GAGCTCTCGGAGTC AACAGG-3'
FLC	At5G10140.1	PDS3110 5'- CGAACTCATGTTGA AGCTTGTT-3'	PDS3111 5'- GGAGAGTCACCGGA AGATTG-3'
AGL24	At4G24540.1	PDS3108 5'- CGGAATTGGTGGAT GAGAAT-3'	PDS3109 5'- CAGGGAAGTGTCGG AGTCAT-3'
SOC1	At2G45660.1	PDS3102 5'- AGCTGCAGAAAACG AGAAGC-3'	PDS3103 5'- TGAAGAACAAGGTA ACCCAATG-3'
LFY	At5G61850.1	PDS4778 5'- ATTGGTTCAAGCAC CACCTC-3'	PDS4779 5'- ACGGACCGAATAGT CCCTCT-3'
FT	At1G65480.1	PDS4706 5'- CTGGAACAACCTTT GGCAAT-3'	PDS4707 5'- AGCCACTCTCCCTC TGACAA-3'
FD	At4G35900.1	PDS4758 5'- CACCTCCTGCAACT GTTCTG-3'	PDS4759 5'- AGCCTCGAAAGAGG TGTTGA-3'
AP1	At1G69120.1	PDS3074 5'- TAGGGCTCAACAGG AGCAGT-3'	PDS3075 5'- CAGCCAAGGTTGCA GTTGTA-3'
Ref. gene ***YLS8***	At5G08290.1	PDS4009 5'-TTACTGTTTCGGTT GTTCTCCATTT- 3'	PDS4010 5'- CACTGAATCATGTT CGAAGCAAGT-3'

Supplemental Table 5: Information used to set up the model simulations for the knockdown mutants.

	k_{mut} [a]	Reference
soc1-2	0.3	[61]
soc1-6	0.25	[61]
agl24-2	2.0	[38]
agl24-1	1.0	[61]
svp-31	0.067	[34]
svp-32	0.033	[34]
svp-41	0.025	[34]

[a] To simulate gene expression in mutants, the expression associated to a knockdown mutant was set to k_{mut}; the value of k_{mut} was adjusted to a fraction of the expression of i observed in the first time-point from wild type Col-0.

References Supplemental Material

1. Srikanth A, Schmid M (2011) Regulation of flowering time: all roads lead to Rome. Cell Mol Life Sci 68: 2013-2037.
2. Andres F, Coupland G (2012) The genetic basis of flowering responses to seasonal cues. Nat Rev Genet 13: 627-639.
3. Chew YH, Smith RW, Jones HJ, Seaton DD, Grima R, et al. (2014) Mathematical models light up plant signaling. Plant Cell 26: 5-20.
4. Gould PD, Ugarte N, Domijan M, Costa M, Foreman J, et al. (2013) Network balance via CRY signalling controls the Arabidopsis circadian clock over ambient temperatures. Mol Syst Biol 9: 650.
5. Keily J, Macgregor DR, Smith RW, Millar AJ, Halliday KJ, et al. (2013) Model selection reveals control of cold signalling by evening-phased components of the plant circadian clock. Plant J.
6. Pokhilko A, Fernandez AP, Edwards KD, Southern MM, Halliday KJ, et al. (2012) The clock gene circuit in Arabidopsis includes a repressilator with additional feedback loops. Mol Syst Biol 8: 574.
7. Schmal C, Reimann P, Staiger D (2013) A Circadian Clock-Regulated Toggle Switch Explains AtGRP7 and AtGRP8 Oscillations in Arabidopsis thaliana. PLoS Comput Biol 9.
8. Jonsson H, Heisler MG, Shapiro BE, Meyerowitz EM, Mjolsness E (2006) An auxin-driven polarized transport model for phyllotaxis. Proc Natl Acad Sci U S A 103: 1633-1638.
9. van Mourik S, Kaufmann K, van Dijk AD, Angenent GC, Merks RM, et al. (2012) Simulation of organ patterning on the floral meristem using a polar auxin transport model. PLoS One 7: e28762.
10. Peret B, Middleton AM, French AP, Larrieu A, Bishopp A, et al. (2013) Sequential induction of auxin efflux and influx carriers regulates lateral root emergence. Mol Syst Biol 9: 699.
11. Grieneisen VA, Xu J, Maree AF, Hogeweg P, Scheres B (2007) Auxin transport is sufficient to generate a maximum and gradient guiding root growth. Nature 449: 1008-1013.
12. Salazar JD, Saithong T, Brown PE, Foreman J, Locke JC, et al. (2009) Prediction of photoperiodic regulators from quantitative gene circuit models. Cell 139: 1170-1179.
13. Song YH, Smith RW, To BJ, Millar AJ, Imaizumi T (2012) FKF1 conveys timing information for CONSTANS stabilization in photoperiodic flowering. Science 336: 1045-1049.
14. van Mourik S, van Dijk AD, de Gee M, Immink RG, Kaufmann K, et al. (2010) Continuous-time modeling of cell fate determination in Arabidopsis flowers. BMC Syst Biol 4: 101.
15. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER (2004) A gene regulatory network model for cell-fate determination during Arabidopsis thaliana flower development that is robust and recovers experimental gene expression profiles. Plant Cell 16: 2923-2939.
16. Mendoza L, Alvarez-Buylla ER (1998) Dynamics of the genetic regulatory network for Arabidopsis thaliana flower morphogenesis. J Theor Biol 193: 307-319.
17. Sanchez-Corrales YE, Alvarez-Buylla ER, Mendoza L (2010) The Arabidopsis thaliana flower organ specification gene regulatory network determines a robust differentiation process. J Theor Biol 264: 971-983.
18. Jung C, Muller AE (2009) Flowering time control and applications in plant breeding. Trends Plant Sci 14: 563-573.
19. Dong Z, Danilevskaya O, Abadie T, Messina C, Coles N, et al. (2012) A gene regulatory network model for floral transition of the shoot apex in maize and its dynamic modeling. PLoS One 7: e43450.
20. Jaeger KE, Pullen N, Lamzin S, Morris RJ, Wigge PA (2013) Interlocking feedback loops govern the dynamic behavior of the floral transition in Arabidopsis. Plant Cell 25: 820-833.
21. Satake A, Kawagoe T, Saburi Y, Chiba Y, Sakurai G, et al. (2013) Forecasting flowering phenology under climate warming by modelling the regulatory dynamics of flowering-time genes. Nat Commun 4: 2303.
22. Welch SM, Roe JL, Dong ZS (2003) A genetic neural network model of flowering time control in Arabidopsis thaliana. Agronomy Journal 95: 71-81.

23. Pose D, Yant L, Schmid M (2012) The end of innocence: flowering networks explode in complexity. Curr Opin Plant Biol 15: 45-50.
24. Jang S, Torti S, Coupland G (2009) Genetic and spatial interactions between FT, TSF and SVP during the early stages of floral induction in Arabidopsis. Plant J 60: 614-625.
25. Helliwell CA, Wood CC, Robertson M, James Peacock W, Dennis ES (2006) The Arabidopsis FLC protein interacts directly in vivo with SOC1 and FT chromatin and is part of a high-molecular-weight protein complex. Plant J 46: 183-192.
26. Li D, Liu C, Shen L, Wu Y, Chen H, et al. (2008) A repressor complex governs the integration of flowering signals in Arabidopsis. Dev Cell 15: 110-120.
27. Tao Z, Shen L, Liu C, Liu L, Yan Y, et al. (2012) Genome-wide identification of SOC1 and SVP targets during the floral transition in Arabidopsis. Plant J 70: 549-561.
28. Mathieu J, Warthmann N, Kuttner F, Schmid M (2007) Export of FT protein from phloem companion cells is sufficient for floral induction in Arabidopsis. Curr Biol 17: 1055-1060.
29. Corbesier L, Vincent C, Jang S, Fornara F, Fan Q, et al. (2007) FT protein movement contributes to long-distance signaling in floral induction of Arabidopsis. Science 316: 1030-1033.
30. Abe M, Kobayashi Y, Yamamoto S, Daimon Y, Yamaguchi A, et al. (2005) FD, a bZIP protein mediating signals from the floral pathway integrator FT at the shoot apex. Science 309: 1052-1056.
31. Wigge PA (2013) Ambient temperature signalling in plants. Curr Opin Plant Biol.
32. Yoo SK, Chung KS, Kim J, Lee JH, Hong SM, et al. (2005) CONSTANS activates SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 through FLOWERING LOCUS T to promote flowering in Arabidopsis. Plant Physiol 139: 770-778.
33. Wigge PA, Kim MC, Jaeger KE, Busch W, Schmid M, et al. (2005) Integration of spatial and temporal information during floral induction in Arabidopsis. Science 309: 1056-1059.
34. Hartmann U, Hohmann S, Nettesheim K, Wisman E, Saedler H, et al. (2000) Molecular cloning of SVP: a negative regulator of the floral transition in Arabidopsis. Plant J 21: 351-360.
35. Gregis V, Andres F, Sessa A, Guerra RF, Simonini S, et al. (2013) Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in Arabidopsis. Genome Biol 14: R56.
36. Deng W, Ying H, Helliwell CA, Taylor JM, Peacock WJ, et al. (2011) FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of Arabidopsis. Proc Natl Acad Sci U S A 108: 6680-6685.
37. Immink RG, Pose D, Ferrario S, Ott F, Kaufmann K, et al. (2012) Characterization of SOC1's central role in flowering by the identification of its upstream and downstream regulators. Plant Physiol 160: 433-449.
38. Michaels SD, Ditta G, Gustafson-Brown C, Pelaz S, Yanofsky M, et al. (2003) AGL24 acts as a promoter of flowering in Arabidopsis and is positively regulated by vernalization. Plant J 33: 867-874.
39. Lee J, Oh M, Park H, Lee I (2008) SOC1 translocated to the nucleus by interaction with AGL24 directly regulates leafy. Plant J 55: 832-843.
40. Wagner D, Sablowski RW, Meyerowitz EM (1999) Transcriptional activation of APETALA1 by LEAFY. Science 285: 582-584.
41. Kaufmann K, Wellmer F, Muino JM, Ferrier T, Wuest SE, et al. (2010) Orchestration of floral initiation by APETALA1. Science 328: 85-89.
42. Hempel FD, Weigel D, Mandel MA, Ditta G, Zambryski PC, et al. (1997) Floral determination and expression of floral regulatory genes in Arabidopsis. Development 124: 3845-3853.
43. Giakountis A, Coupland G (2008) Phloem transport of flowering signals. Curr Opin Plant Biol 11: 687-694.
44. Torti S, Fornara F, Vincent C, Andres F, Nordstrom K, et al. (2012) Analysis of the Arabidopsis shoot meristem transcriptome during floral transition identifies distinct regulatory patterns and a leucine-rich repeat protein that promotes flowering. Plant Cell 24: 444-462.

45. Michaels SD, Amasino RM (1999) FLOWERING LOCUS C encodes a novel MADS domain protein that acts as a repressor of flowering. Plant Cell 11: 949-956.
46. van Leeuwen W, Ruttink T, Borst-Vrenssen AW, van der Plas LH, van der Krol AR (2001) Characterization of position-induced spatial and temporal regulation of transgene promoter activity in plants. J Exp Bot 52: 949-959.
47. Hames C, Ptchelkine D, Grimm C, Thevenon E, Moyroud E, et al. (2008) Structural basis for LEAFY floral switch function and similarity with helix-turn-helix proteins. EMBO J 27: 2628-2637.
48. de Folter S, Immink RG, Kieffer M, Parenicova L, Henz SR, et al. (2005) Comprehensive interaction map of the Arabidopsis MADS Box transcription factors. Plant Cell 17: 1424-1433.
49. Wilczek AM, Roe JL, Knapp MC, Cooper MD, Lopez-Gallego C, et al. (2009) Effects of genetic perturbation on seasonal life history plasticity. Science 323: 930-934.
50. Lee JH, Ryu HS, Chung KS, Pose D, Kim S, et al. (2013) Regulation of temperature-responsive flowering by MADS-box transcription factor repressors. Science 342: 628-632.
51. Immink RG, Tonaco IA, de Folter S, Shchennikova A, van Dijk AD, et al. (2009) SEPALLATA3: the 'glue' for MADS box transcription factor complex formation. Genome Biol 10: R24.
52. Pose D, Verhage L, Ott F, Yant L, Mathieu J, et al. (2013) Temperature-dependent regulation of flowering by antagonistic FLM variants. Nature.
53. Lloyd J, Meinke D (2012) A comprehensive dataset of genes with a loss-of-function mutant phenotype in Arabidopsis. Plant Physiol 158: 1115-1129.
54. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, et al. (2003) Global analysis of protein expression in yeast. Nature 425: 737-741.
55. Schmid M, Uhlenhaut NH, Godard F, Demar M, Bressan R, et al. (2003) Dissection of floral induction pathways using global expression analysis. Development 130: 6001-6012.
56. Mathieu J, Yant LJ, Murdter F, Kuttner F, Schmid M (2009) Repression of flowering by the miR172 target SMZ. PLoS Biol 7: e1000148.
57. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80.
58. Taoka K, Ohki I, Tsuji H, Furuita K, Hayashi K, et al. (2011) 14-3-3 proteins act as intracellular receptors for rice Hd3a florigen. Nature 476: 332-335.
59. Yu H, Xu Y, Tan EL, Kumar PP (2002) AGAMOUS-LIKE 24, a dosage-dependent mediator of the flowering signals. Proc Natl Acad Sci U S A 99: 16336-16341.
60. Lee JH, Yoo SJ, Park SH, Hwang I, Lee JS, et al. (2007) Role of SVP in the control of flowering time by ambient temperature in Arabidopsis. Genes Dev 21: 397-402.
61. Liu C, Chen H, Er HL, Soo HM, Kumar PP, et al. (2008) Direct interaction of AGL24 and SOC1 integrates flowering signals in Arabidopsis. Development 135: 1481-1491.
62. Moon J, Lee H, Kim M, Lee I (2005) Analysis of flowering pathway integrators in Arabidopsis. Plant Cell Physiol 46: 292-299.
63. Cavelier G, Anastassiou D (2004) Data-based model and parameter evaluation in dynamic transcriptional regulatory networks. Proteins 55: 339-350.
64. Buchler NE, Louis M (2008) Molecular titration and ultrasensitivity in regulatory networks. J Mol Biol 384: 1106-1119.

Chapter 3

Predicting the effect of SNPs on flowering time

Felipe Leal Valentim, Gerco C. Angenent and Aalt D.J. van Dijk

In preparation for submission

Abstract

Transcriptional regulation plays a critical role in developmental evolution, but it is currently unclear which particular regulatory elements are involved in the evolution of flowering time control. Here, we present a computational approach to identify *cis*-regulatory single nucleotide polymorphisms (SNPs) that have an effect on flowering time in *Arabidopsis thaliana*. Starting with genes that have experimentally been implicated to affect flowering time, we first show that, by using experimentally determined binding sites and/or matches to known binding site motifs, we can identify SNPs that are highly discriminative in the classification of the flowering time phenotype of different Arabidopsis accessions. Subsequently, we interrogated literature to formulate hypotheses for the molecular mechanisms underlying effects of SNPs on gene regulation and the resulting effects on flowering time. The SNPs with strongest association with flowering time, i.e. the most discriminative in the phenotype classification, include variations within the *FIONA 1* (*FIO1*) and *FLOWERING LOCUS T (FT)* genes, in which our method predicts that the SNPs disrupt the binding of MADS-box transcription factors, thus resulting in changes in gene expression and consequently a change in flowering time. Finally, our method reveals statistical dependencies between the selected SNPs. For several cases, hypotheses for the molecular mechanism explaining the flowering time phenotypes are only consistent when this dependency is taken into account.

Introduction

Understanding the molecular basis of adaptation of flowering time control is a major challenge in plant biology. Recent studies suggest that *cis*-regulatory mutations that result in changes in gene expression may play critical roles in the evolution of flowering time control [1]. However to what extent and by which mechanisms these mutations result in phenotypic differences is yet unexplored. In humans, experimental evidence shows that the presence of single nucleotides polymorphisms (SNPs) in regions of binding sites of transcription factor (TFs) may correlate with differences in binding affinity among individuals [2], possibly by mechanistically disrupting the TF binding [3]. Genome-wide association studies (GWAS) have been applied to identify the association of a vast amount of SNPs with several trait phenotypes in plants, including flowering time [4-7]. However, a major hurdle in the interpretation of GWAS results and the understanding of the role of SNPs is that most identified associations point to relatively large regions of correlated variants. This is due to linkage disequilibrium (LD), and makes it difficult to precisely identify the SNPs that are causal for the phenotype [8]. Consequently, it is not possible to understand the molecular mechanism linking the SNP to the phenotype. Furthermore, the vast majority of SNPs identified in GWAS in *Arabidopsis* are located in non-coding regions [4], thus it is likely that most SNPs if they have an effect, have a regulatory role, e.g. by changing gene expression. Based on that, expression quantitative trait loci (eQTLs) can be used to identify the targets that are likely to be affected by SNPs identified in a

GWAS [7], however these studies also point to regions of LD instead to single SNPs. Therefore, methods for associating individual SNPs to a phenotype are necessary.

As developments in technology popularize ChIP-Seq and ChIP-ChIP experiments, approaches based on experimentally determined binding sites have been proposed to annotate a subset of SNPs identified by GWAS. The most popular idea is to filter GWAS results in order to select the SNPs that overlap the experimentally determined binding sites [8,9]. Then, a common assumption (based on experimental evidence [2]) is that a SNP in a binding site leads to differences in TF binding between accessions. Alternatively to using experimentally determined binding sites, computationally determined binding sites can be used [10] with the same assumption. These computational approaches usually rely on assessing the presence of a consensus binding motif in the DNA, which, however, does not necessarily correlate with *in vivo* binding. For this study, we explore the combination of both ideas to identify SNPs that are related to the flowering time phenotype: i.e. both experimentally determined and computationally determined binding sites are used to pinpoint SNPs that affect TF binding and flowering time. We then analyse the SNPs that have the strongest association with flowering time variation to understand the mechanism linking the SNP with the phenotype.

Results

Identifying genes with disruptions in their regulatory region

We aimed at identifying SNPs that have an effect on the transcriptional regulation of the genes involved in flowering time control. The first step was to identify genes whose sequences contain SNPs by comparing sequences of 374 available accessions relative to the Col-0 sequence. We focused primarily on 174 flowering time genes [11] that have experimentally been implicated in the control of flowering time. From the 174 flowering time genes, all of them have SNPs in the genic sequence (gene body plus 2kb upstream sequence); and on average we find 13 SNPs in each gene. We first focused on the SNPs overlapping the position of experimentally determined binding sites of TFs involved in plant reproduction. The positions of these binding sites were determined as the ChIP-seq peaks published by the studies listed in Table 1 (see Material and Methods). We saw that there is at least one ChIP-seq peak of any TF in 171 of the flowering time genes; and on average 32% (sd. 23%) of the genic sequence is covered by ChIP-seq peaks. From these, the average number of accessions with a SNP overlapping a ChIP-seq peak region in any of the flowering time genes is 254 (sd. 99). It can therefore be concluded that having a SNP overlapping a ChIP-seq peak is a common event because of the great percentage of genic sequence covered by peaks; and this event alone may not be indicative for disruption of the TF-DNA binding.

Table 1 - Overview of ChIP-seq studies for TFs involved in plant reproduction used in this work.

Gene		Family	Developmental role	Tissue	References
AGAMOUS	AG	MADS-box transcription factors	Specification of floral-organ identity	Flower buds stage 5	[12]
APETALA1	AP1	MADS-box transcription factors	Specification of floral-organ identity	Inflorescence meristem	[13]
APETALA2	AP2	AP2-like family	Specification of floral-organ identity	Inflorescences	[14]
APETALA3	AP3	MADS-box transcription factors	Specification of floral-organ identity	Flower buds stage 5	[15]
FLOWERING LOCUS C	FLC	MADS-box transcription factors	Control of flowering time	12 days old seedlings	[16]
FLOWERING LOCUS M	FLM	MADS-box transcription factors	Control of flowering time	15 days old seedlings	[17]
LEAFY	LFY		Control of flowering time	Flower buds	[18]
PISTILLATA	PI	MADS-box transcription factors	Specification of floral-organ identity	Flower buds stage 5	[15]
SEPALLATA3	SEP3	MADS-box transcription factors	Specification of floral-organ identity	Inflorescences stage 1-12	[19]
SHORT VEGETATIVE PHASE	SVP	MADS-box transcription factors	Control of flowering time	2 weeks old seedlings	[20]
SHORT VEGETATIVE PHASE	SVP	MADS-box transcription factors	Control of flowering time	Inflorescences stage 1-11	[20]
SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1	SOC1	MADS-box transcription factors	Control of flowering time	Transition apices	[21]

Nine of the eleven TFs whose binding sites have been considered (Table 1) in the analysis belong to the MADS-domain protein family. Members of this protein family are known to bind to CArG-box motifs [22]. Because of this, we subsequently focused on the set of SNPs overlapping the position of CArG-boxes within the genic sequence of the selected genes. For that, we searched for the presence of the four most common types of CArG-box motifs in both DNA strands of the Arabidopsis genome (see Material and Methods). On average, 47 (sd. 31) nucleotides of a genic sequence are part of CArG-box motif, and 34 (sd. 63) accessions have a SNP overlapping a CArG-box region per gene; this regardless of the position of the CArG-boxes within the genic region. We also focused on the subset of SNPs that overlap CArG-boxes that are located within ChIP-seq peaks. From the 174 flowering time genes, in 171 and 170 we find at least one ChIP-seq peak and at least one CArG-box motif, respectively. From these, there are 114 for which we observe at least one CArG-box overlapping a ChIP-seq peak. From these, 70 genes have a SNP disrupting the CArG-box that is also located within a peak. For these 70 genes, on average there are 22 (sd. 36) accessions with such SNPs.

We noted that the number of genes targeted by each of the TFs varies considerably (Table 2); e.g. whilst 142 flowering time genes are targeted by SHORT VEGETATIVE PHASE (SVP), only 9 are targeted by SUPRESSOR OF OVEREXPRESSION OF CONSTANS1 (SOC1). One reason for this could be that the total number of peaks also varied considerably among the ChIP-seq studies

considered in this work. We note that for this analysis we included both, TFs involved on the flowering time regulation (AP2, AP3, FLC, FLM, LFY, SEP3, SVP, SOC1) but also on the floral organ specification (AG, AP1 and PI). A future step will be to re-generate the results separating both sets of TFs.

Table 2 – Number of flowering time genes targeted by each TFs involved in plant reproduction used in this work.

TF	Number of targeted flowering time genes	Number of genes with a SNP overlapping a CArG-box within a ChIP-seq peak
AG	36	23
AP1	98	48
AP2	30	18
AP3	35	24
FLC	16	9
FLM	135	58
LFY	20	8
PI	41	25
SEP3	26	15
SVP	142	61
SOC1	9	0

Phenotypic variance predicted by SNPs

Based on the previous analyses, three sets of SNPs were listed: 1) SNPs overlapping a ChIP-seq peak within the genic region; 2) SNPs overlapping a CArG-box motif within the genic region; or 3) SNPs overlapping a CArG-box that overlaps a ChIP-seq peak within the genic region. We assume that these SNPs potentially have an effect on the regulatory network of flowering time control by disrupting the binding of the TF. In order to model the relationship between SNPs and the flowering time phenotype, we developed predictive models based on each of the mentioned sets of SNPs. We compared the three models by assessing their ability to predict the flowering time phenotype given the information about SNPs.

To do that, decision tree models were fitted and used to predict the flowering phenotypes. The idea here is that the model which best predicts the flowering time is the one using the most explanatory set of SNPs. As input, we created a table that indicates whether a SNP is present or not in each of the flowering time genes, this for each accession. The flowering time phenotypes were binary categorized as “early” or “late” based on comparison against the flowering time observed for the reference Col-0 (see Table S1). Thus, our decision tree model was completely binary; the input variables are represented by “presence” or “absence” of SNPs in a given gene for each accession, and the output variable indicates whether an accession is “early” or “late” flowering. The decision tree models then search for combinations of SNPs that together predict whether an accession is “early” or “late” flowering. To assess the models, we defined two measures that quantify the quality of the predictions; “Precision” and “Recall” (see Materials and Methods). Hereafter, the decision tree models are referred

as ‘Model 1’, ‘Model 2’ and ‘Model 3’, corresponding respectively to the model using SNPs overlapping a ChIP-seq peak within the genic region, the model using SNPs overlapping a CArG-box motif within the genic region, and the model using SNPs overlapping a CArG-box that overlaps a ChIP-seq peak within the genic region.

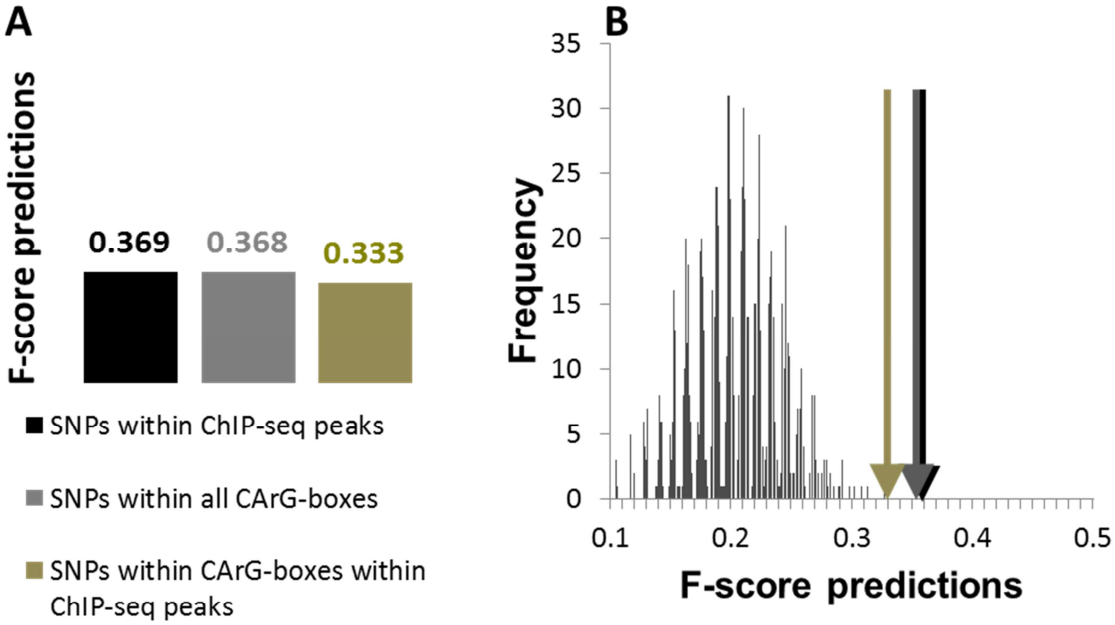


Figure 1: Assessment of overall performance of the predictions using different sets of SNPs. (A) F-score values obtained by following a leave-one-out strategy. For each of the 374 accessions used in this work, we fitted the decision tree models using 373 accessions and predicted the phenotype for the one not used in the fitting. The F-score represents the $F_{\beta=1}$ between of Precision and Recall. **(B)** Frequency distribution of F-scores from permutation tests over each of the models. For this test, 1000 sets of randomly shuffled flowering time phenotypes were generated and compared against the predicted flowering time phenotypes.

The decision tree model using SNPs overlapping a CArG-box that overlaps a ChIP-seq peak (Model 3) uses the smallest subset of SNPs, which supposedly are most relevant. Before analysing this model in detail, it is important to verify that its performance in predicting flowering time is comparable to that of the alternative models (Model 1 and 2) which use larger sets of SNPs as input. The value for Precision is highest in Model 3 (0.46, 0.52 and 0.56 for Models 1, 2 and 3; respectively) whilst the value for Recall is highest in Model 1 (0.31, 0.28 and 0.24 for Models 1, 2 and 3; respectively). We assessed the model also by calculating an F1-score as overall performance measure (see Material and Methods). We observe only a small difference between the F1-score of the three models (Model 1 = 0.37, Model 2= 0.37, Model 3= 0.33). Quantitative comparison among the models is inconclusive (Figure 1); however, because of the biological relevance of the SNPs included in Model 3 and because Model 3 shows the greatest value of Precision, we focus subsequent analysis on this model and only present results for this model. A graphical representation of the decision tree is shown in Figure 2. One

potential concern of the type of analysis that we present is that predictive performance might be highly related to population structure. Hence, we checked for population structure within the SNP data. Heat map clustering analysis revealed no structure (Figure 3).

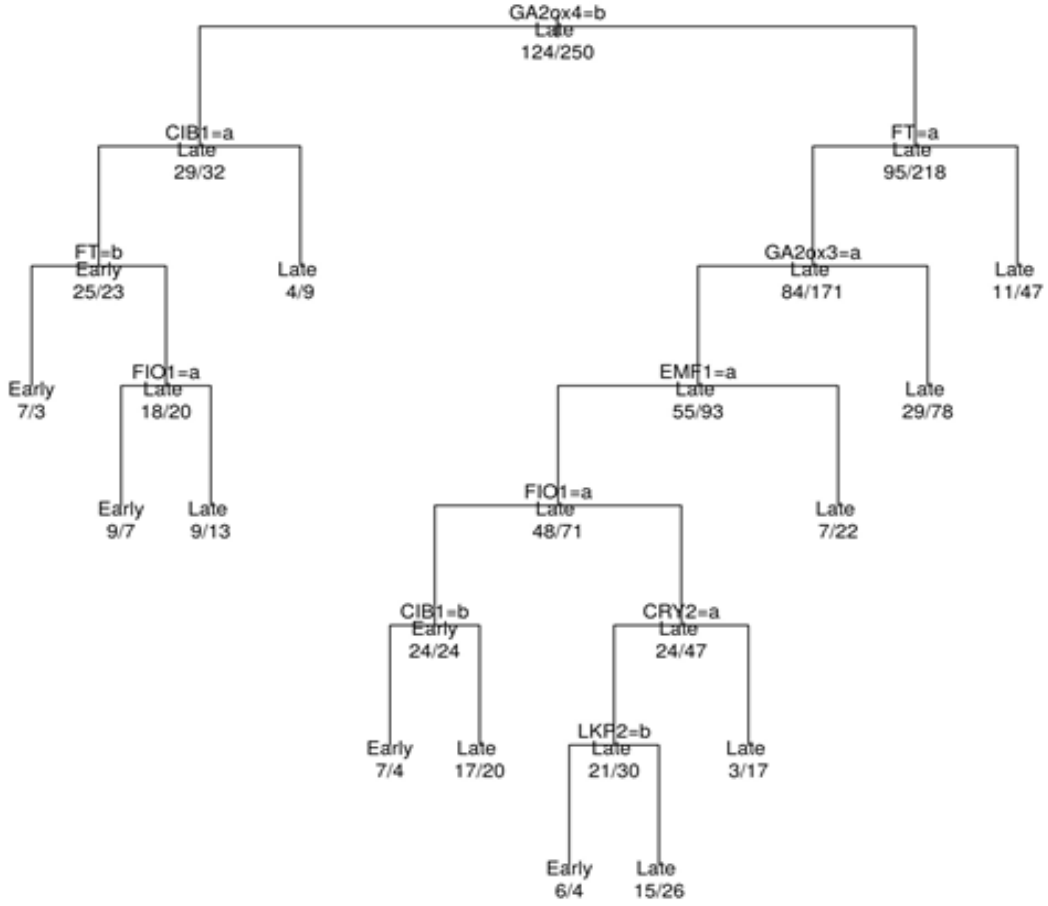


Figure 2: Graphical representation of the decision tree for the 'Model 3'. The tree shows the relationship between the variables. The gene represented by each node of the tree is equalized to either characters b or a, indicating respectively the scenarios whether that gene has or does not have the SNP disrupting the binding site (SNP "present" or "absent", represented by the letter "b" and "a", respectively); e.g. GA2OX4=b. The numbers below the gene name in each node represent the number of accessions that show "Early"/"Late" flowering phenotype; e.g. at the root of the tree, the numbers 124/250 indicate that before any SNP is taken into account, 124 accessions show "Early" flowering phenotype, while 250 accessions show "Late" phenotype. Based on these numbers, the tree predicts that the accessions have "Late" flowering phenotype (phenotype indicated below the gene names). If the scenario indicated by the node is met (e.g. GA2OX4=b, indicating the scenario where the accession has the SNP disrupting a binding site in GA2OX4) the tree processes the branches of the left side to take the decision, otherwise it processes the branches of the right side. For example, when there is a SNP disrupting a binding site in GA2OX4 (GA2OX4=b); then the tree analyses the node for the gene CIB1. The numbers below the node of the gene CIB1indicate that in total 61 accessions have the SNP disrupting a binding site in GA2OX4; from which 29 accessions show "Early" phenotype, and 32 show "Late" phenotype. Based on these numbers, the tree again predicts that the accessions have "Late" flowering phenotype (indicated below the gene name). This same reasoning is applied until the process reaches the leaves of the tree; when the final decision is actually made. This decision is based on the numbers of accessions that show "Early"/ "Late"

flowering phenotype. For example, if the leaf node shows the numbers 7/4, it means that the accession that reached that node will be classified as “Early” because the observed phenotype of 7 out of 11 accessions is “Early” flowering.

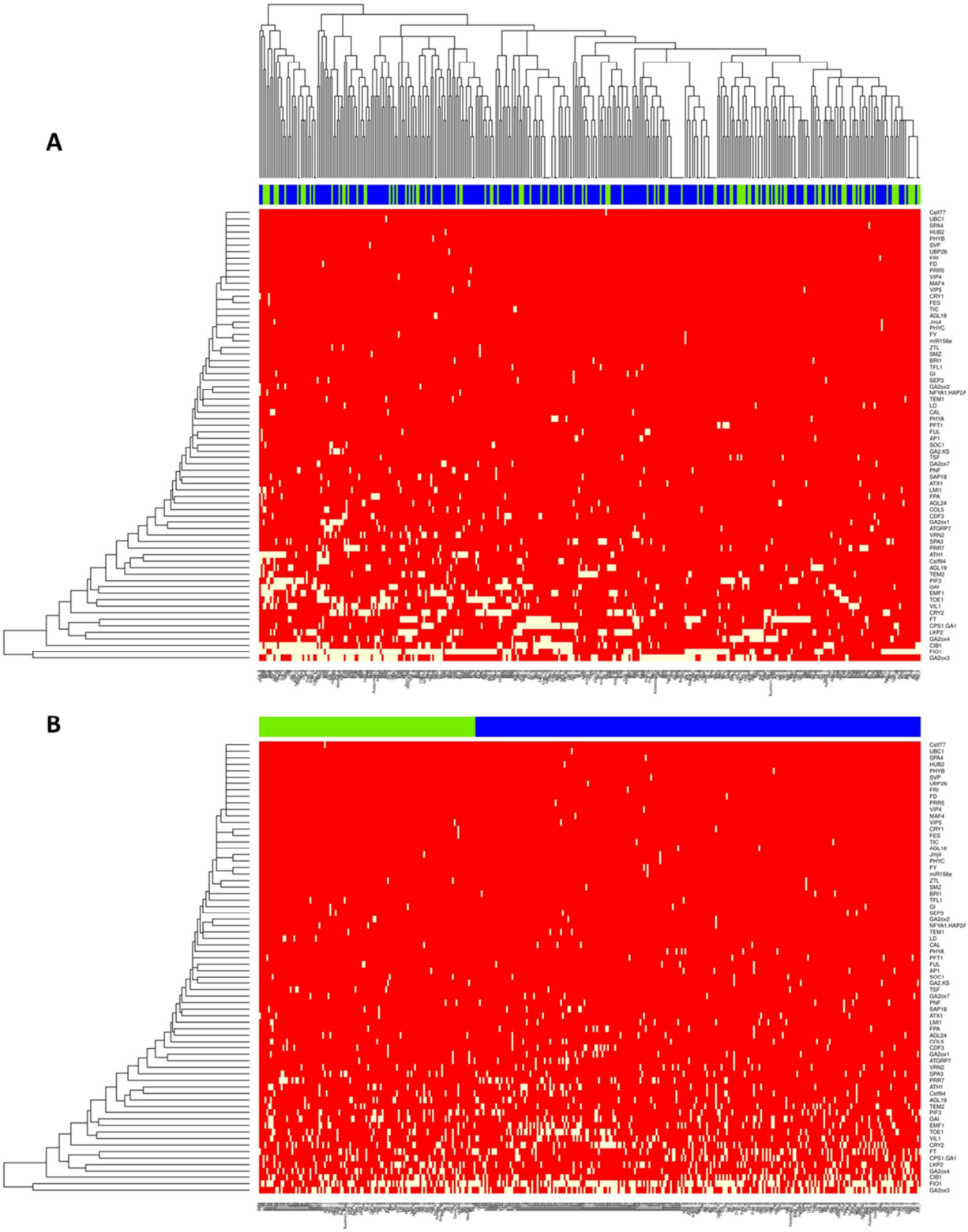


Figure 3: Heat maps of SNPs within the flowering time genes of each accession. The rows indicate the flowering time genes and the columns indicate the accessions. A white square indicates that there is at least one SNP overlapping a CArG-box within a ChIP-seq peak within the genic region (gene plus promoter). In **(A)**, the

accessions are clustered based on their SNPs, and in **(B)** the accessions are grouped based on their phenotype. The blue and green colours on top of heat maps indicate the flowering time phenotype of the accessions; where green indicates "Early" flowering and blue indicates "Late" flowering in relation to the flowering time of Col0. If there were population structure in this SNP data set, the clustering analysis would have grouped the accessions according to their phenotypes.

Identifying SNPs with effect on the regulatory network of flowering time control

The goal is to use variable importance ranking in the decision tree model to derive a set of highly discriminative genes in the classification of the flowering time phenotypes. In total, 16 variables were ranked as important, i.e. 16 genes in which SNPs in their binding sites influence flowering time according to the model (Table 3). Before examining the gene selection, we first evaluated the significance of the variable importance ranking analysis. For that, we assessed if the method is able to distinguish flowering time genes from non-flowering time genes; this based on variable ranking analysis when using SNP data of random sets of genes added to the flowering time genes (See Material and Methods). The random sets of genes were created by adding 174 genes random non-related to flowering time to the list of 174 flowering time genes. We observed that on average 93.7% of the genes ranked as important by the variable selection are indeed flowering time genes. We used the models fitted with each of the random sets of to predict the flowering time phenotypes. From the predictions we observed F-scores value of 0.37, 0.37 and 0. 0.33; similar to that observed in the original models.

We next assessed if the genes ranked as the most important have the largest influence on phenotype prediction quality (See Material and Methods). We observed that when the genes ranked as the most important variable (FIO1) or the less important variable (GAI) are separately removed from the model, the F-score value of the phenotype predictions does not change compared to that of the original model. However, when we removed from the input table all the genes selected as important and we fit a new model, the method was not able to reconstruct a meaningful tree (F-score=0). Overall, we conclude that the method selects a unique set of genes, and that individual genes contribute the same for quality of the predictions regardless of the variable ranking position.

Effect of SNPs on the regulatory network of flowering time control

We identified and characterized the SNPs that are causing the disruptions in the binding sites. For that, we examined the 16 selected genes to identify the specific SNPs that are affecting the binding sites (see columns 1 and 2 of Table 3 for the names of the selected genes and the values obtained from the variable importance ranking analysis, respectively). Based on the above described decision tree 'Model 3', we inferred the role of the SNPs in the flowering time, i.e. we determined if the SNPs is predicted to confer "Late" or "Early" flowering phenotype to the accessions (see column 7 of Table 3). For that, we calculated the percentage of accessions that show "Late" flowering phenotype when there

is a SNP affecting the binding site of the selected gene and compared with the percentage of accessions that show "Late" flowering phenotype when there is no binding site disruption. This is done by using the accessions as grouped by each split of the tree. For instance, for the variable in the root of the tree GA2ox4, 61 accessions have the SNP disrupting the binding site (GA2ox4=b) and 313 accessions don't carry the disruptive SNP (GA2ox4=a). From these, 69% (218 out of 313) and 52% (32 out of 61) show "Late" flowering phenotype, for GA2ox4=a and GA2ox4=b, respectively (see column 6 of Table 3). Based on these numbers, we infer that the SNP is involved in conferring an "Early" flowering phenotype. The rationale behind this is that when the disruptive SNP is observed (GA2ox4=b), then the percentage of accessions with "Late" phenotype is decreased compared to GA2ox4=a. For a few cases, the gene is used in more than one split of the tree. For these, we used the split in which more accessions are involved. For another few cases, the gene is ranked as important but not represented by the tree. This is because there may be candidate variables that are important but are not used in a split. In such a case, the top competing variables are also tabulated at each split [23]. For these cases, the roles of the disruptive SNPs were determined based on the Early/Late flowering phenotypes of accessions with and without the SNP, irrespective of the decision tree.

We compared those decision tree model-based predictions with the expected impact of SNPs from literature evidence, as follows. We first determined the regulatory relationship between the TF whose binding site is being disrupted and the targeted flowering time gene (see column 5 of Table 3). This is done using co-expression analysis and predicts if a TF may activate or supress a target gene (see Table S1). We realise that co-expression correlation only gives an indication about the transcriptional relationship between TF and its targets, but unfortunately, more detailed experimental evidence is often lacking. In addition, we define, based on knock-out and overexpression studies, the role of the selected flowering time genes on the control of flowering time, i.e. if they are inducers or repressors of floral transition (see columns 9 and 10 of Table 3). From these, we infer whether a disruption of a TF binding would result in increase or decrease in expression of the target gene and we infer the expected effect of this change in gene expression on flowering time (see column 8 of Table 3). We then assessed if this evidence-driven inference is consistent with the role statistically inferred for the SNPs by the decision tree model. The results are summarized in Table 3 and some examples are further discussed below.

Knock-out mutants of *FIO1* show early flowering phenotypes [24], i.e. *FIO1* is a repressor of flowering. This gene is ranked as the most important variable, which means that there are SNPs in its genic region that are highly discriminative in the classification of the flowering time phenotype. More specifically, there is one single SNP within the *FIO1* genic region that overlaps the position of a CArG-box within the binding site of the TF SVP. Thus, we hypothesize that this SNP interferes with the regulation of *FIO1* by the TF SVP. Further, co-expression analysis shows that there is a negative correlation between the expression of *FIO1* and *SVP, i.e*. it is likely that SVP is a repressor of *FIO1*

expression. Therefore, by disrupting this binding site of SVP, this TF would not be able to repress the expression of *FIO1* and consequently, it would result in an increase in *FIO1* gene expression and a consequent delay in flowering time. By analysing the tree, we observed that when there is not a SNP disrupting the binding site of *FIO1* (FIO1=a), 50% of the accessions shows a "Late" flowering phenotype. In contrast, when there is a SNP disrupting the binding site of *FIO1* (FIO1=b), then the proportion of accessions that show a "Late" flowering phenotype is increased to 66%. Based on this information, we infer that the specific SNP is likely to confer "Late" phenotype by disrupting a binding site of the regulator(s) of *FIO1*. This is consistent with the data-driven inferred SNP role.

Another gene ranked among the most important variables is the florigen *FT*. This gene is regulated in the leaves and its encoded protein moves to the meristem to activate the floral transition [25]. In total, 348 accessions have SNPs over its genic region. From these, 233 accessions have SNPs that overlap one of the ChIP-seq peaks in the genic region; and 99 accessions have SNPs that overlap one of the CArG-boxes. More interestingly, only 75 accessions have a SNP that overlaps the single CArG-box that is located within the ChIP-seq peaks (see Figure 4). By analysing the tree, we observe that when there is a SNP disrupting the binding site in *FT* (*FT*=b), 81% of the accessions (47 out of 58) show "Late" flowering phenotype. In contrast, when there is not a SNP disrupting the binding site of FT (*FT*=a), then the proportion of accessions that show "Late" flowering phenotype is reduced to 67% (171 out of 255). Based on this information, we infer that the specific SNP is likely to confer a "Late" phenotype by affecting a binding site of the regulators in *FT*. This SNP disrupts a CArG-box that is bound by three MADS-domain TFs: AP1, FLC and FLM (see Figure 4). From these, AP1 is acting in the floral meristem after the switch to reproductive development, while FLC and FLM play a role in the control of flowering time [26]. Therefore, we further investigated the relationship between FLC, FLM and their target *FT*. Co-expression analysis suggests that FLM is an activator of the expression of *FT*, whilst FLC is a repressor. However, according to the recent study of Posé et al. [17], two *FLM* splicing forms are active, the ß-form which is able to bind DNA and acts in combination with SVP as suppressor, and FLMδ that is not able to bind DNA and acts as a competitor of FLMß. The hypothesis in this case would be that the specified SNP disrupts the binding of the two MADS-box TFs, FLMδ and FLC. Because the analysis of our decision tree model predicts that the SNP confers "Late" flowering phenotype to the accessions, we hypothesize that is likely that there is a reduction in *FT* expression as result of the mutation. Based on this observation and the information about *FLM* and *FLC*, our prediction is that the SNP does not interfere with binding of FLMß or FLC, which are repressors of *FT*; but it might interfere with the binding of other known factors that are not represented in our dataset.

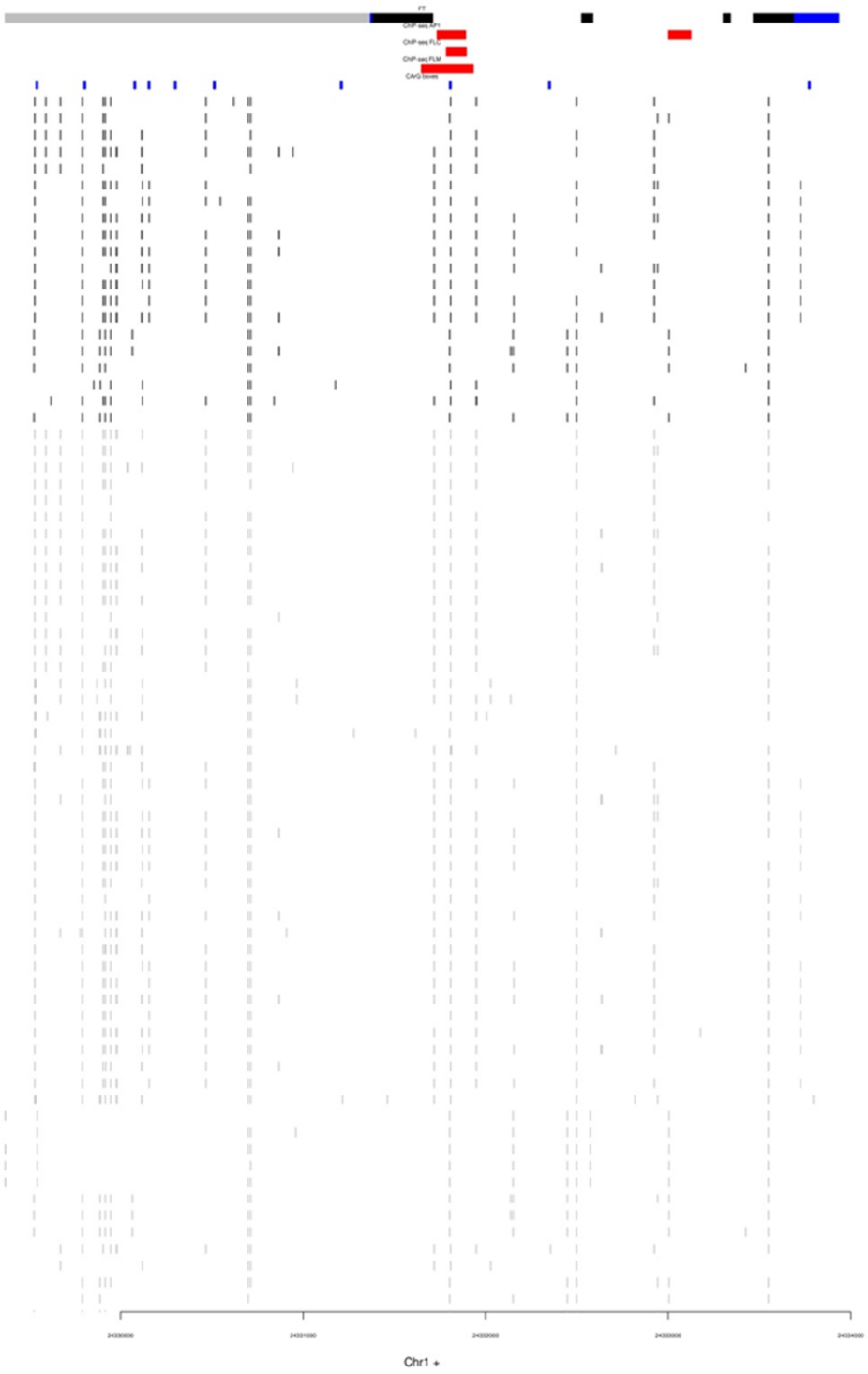


Figure 4: Graphical representation of the SNPs over the genic region of the gene FT. The gene structure is represented by the bar on top of the figure, where black rectangles delimitate the region of the exons, and grey rectangles delimitate the 2kb upstream the gene region. The positions of ChIP-seq peaks of the 11 TFs involved in plant development are indicated by the red columns. The positions of the CArG-boxes are shown by the blue columns. The SNPs of different accessions are represented by the black or grey columns, representing accessions

that show “Early” or “Late” flowering time respectively. Only accessions which have a SNP overlapping a CArG-box that is located within a ChIP-seq peak are represented.

Two *GA2OX* genes are among the selected genes. *GA2OX* genes are thought to be repressors of flowering but single knock-out mutants show no flowering phenotype [27]. We identified one SNP within *GA2OX4* that is possibly disrupting the binding site of three MADS-box TFs: AP1, PI and AG. These TFs only act after the floral transition. Hence, based on available information alone is difficult to speculate when a SNP disrupting that binding site would result in an “Early” or “Late” flowering phenotype. Based on the tree however, for this gene there is a decrease in the proportion of accessions with a “Late” flowering phenotype when we observe a SNP disrupting the binding site. Therefore, we infer the SNP is likely to confer an “Early” phenotype to the accession.

Another *GA2OX* gene ranked as important is *GA2OX3*. For this gene, we predict that a SNP in its regulatory region is potentially disrupting the binding site of the flowering time regulator SVP – which is a repressor of *GA2OX3*, according to co-expression analysis. By analysing the tree we infer that this SNP is responsible to confer “Late” flowering phenotype to the accessions. Consistently, based on evidences, we would expect that a disruption in the binding of SVP in *GA2OX3* would result in an increase of gene expression and a consequent delay in flowering time.

LKP2 overexpression leads to late flowering [28], which indicates that this gene represses the floral transition. We identified a SNP that disrupts the binding of SVP in the *LKP2* regulatory region. Furthermore, we identified that SVP acts as an activator of *LKP2* expression according to co-expression analysis. Thus, we expect to observe a reduction in *LKP2* if the binding of SVP is disrupted; and as consequently early flowering time phenotype. This is consistent with the prediction by the decision tree model (Table 3).

For *TOE1* and *CRY2*, the analysis predicts that disruptive SNPs confer respectively “Early” and “Late” flowering phenotypes to the accessions. For both genes, this prediction was based on the binding site of AP1. Because AP1 only acts after the floral transition, the precise mechanism by which this SNP may confer different flowering time phenotypes remains unclear.

Discussion

In this study, we analysed the effect of SNPs on the gene regulatory network of flowering time control. By focusing on SNPs that overlap the position of possible regulatory motifs in the flowering time genes, we developed models that predict the flowering time phenotype. We identified the SNPs in regulatory regions of 16 genes that are discriminative in the classification of the flowering time phenotypes of the Arabidopsis accessions. We show that the model predictions and SNPs selection are significantly better than random, and we investigate the effect of each identified SNPs on gene regulation and the resultant effect on the flowering time.

It is expected that the great majority of the regulatory SNPs is to be found within the regions delimited by experimentally determined binding regions [29]. However, partially because of the low resolution of ChIP-seq data, we observed that on average 1/3 of the sequence of a flowering time gene is covered by ChIP-seq peaks. Thus, regulatory SNPs are difficult to pinpoint among the sea of polymorphisms localized within binding regions determined by ChIP-seq studies. To overcome this issue, we focused our analysis on the subset of SNPs that are located within the ChIP-seq peaks and that are part of CArG-box motifs. Note that by focusing on the subset of SNPs that are part of CArG-box motifs we only expect to find regulatory SNPs in targets of MADS-domain proteins and therefore, excluding LFY, AP2 and other TFs from this analysis. A straightforward way to improve our approach would be by focusing, in addition to the CArG-box motifs, also on the subset of SNPs that are part of motifs bound by non-MADS-box TFs – e.g. LFY and AP2 family members.

Regarding the motifs bound by LFY, alignment of its target sequences from SELEX experiments revealed a binding preference for a core 7-bp consensus motif (CCANTG[G/T]) [18]. However, the simple presence or absence of this motif was shown to be a poor predictor of LFY binding. Alternatively, match scores of position-specific scoring matrix (PSSM) [30] for a 19-bp motif (with the same 7-bp consensus core) correlated well with experimentally measured LFY DNA binding affinity. In such a case, we propose that one way to incorporate the subset of SNPs that are part of motifs bound by LFY would be by using the PSSM as described by Moyroud et al. [18], then defining a threshold for the PSSM match score in order to identify regions that are likely to be bound by LFY. Regarding AP2, its binding preference has not been clearly characterized yet. In addition to LFY and AP2, our approach did not yield any target of SOC1. This is because from the 174 flowering time genes, only 9 genes are targeted by SOC1, and from these, none have a SNP that overlaps a CArG-box that is located within a SOC1 ChIP-seq peak.

Previously, genome-wide association studies (GWAS) have been applied to find SNPs that are linked to a particular disease or trait phenotype. For example in [6], SNPs located in the vicinity of a number of genes, *a-priori* associated with different traits (including flowering time under different conditions) were tested for over-representation among Arabidopsis accessions that share similar phenotypes. Our approach focuses on putative *cis*-acting SNPs that are located nearby the *a-priori* identified flowering time genes. The advantage of our more targeted approach is that it is able to detect associations of SNPs to the flowering phenotype even for cases in which the SNP is observed only in a small fraction of the accessions, and thus it is not enriched for all accessions with a certain phenotype. For example, for SPA3 as little as 24 accessions have the SNP identified as having regulatory effect on gene regulation (see Figure S1). In addition our method enables to investigate dependencies between SNPs because the classification tree uses combinations of SNPs as predictors for the phenotype.

In total, 70 flowering time genes have a SNP affecting the consensus CArG-box motif that is located within the experimentally determined binding region (i.e. the ChIP-seq peak). Because of the location of these SNPs, all these 70 cases potentially have an effect on gene regulation. However, our approach selects only 16 cases as having an important association with the flowering time phenotype. This means that certain SNPs are predicted to change the sequence of the CArG-box, but not to change its properties, i.e. affinity to the TF. In order to understand this, it is necessary to examine the specific nature of the mutations and to check how it changes the sequence and structural properties of the CArG-boxes. From our results, we observed no clear sequence pattern that can be used to distinguish a CArG-box polymorphism with an effect on gene regulation from those changes that have no effect. In both cases, SNPs are found over nearly all the positions of the CArG-box and the mutated nucleotides are diverse (see Figure S2). Accordingly, and based on recent studies [31], it can be suggested that there may be structural DNA properties being affected by some of the SNPs pinpointed by our approach.

Our approach is based on predictive decision tree models and the gene selection is based on variable importance ranking analysis. The goal is to identify the smallest possible set of genes (in our case on the basis of their SNPs) that can still achieve good performance for the predictions (in our case of flowering time phenotype). Variable importance evaluation functions can be separated into two groups: those that use the model information and those that do not. We decided for a variable importance evaluation function that uses the tree model information because it selects variables with possible meaningful relationship among them. This approach can be continued by exploring alternative variable selection methods (e.g. elastic net [32]); or by exploring alternative predictive methods (e.g. Random Forest [33]).

Importantly, regarding the classification of the phenotype, there is a statistical relationship between the genes in the same path from the tree root to one of the leaf nodes. Interestingly, these relationships can be explored in order to understand their biological nature. For example, by looking at the tree structure we observed that the gene *GA2OX4* is firstly analysed; and if there is a SNP disrupting its binding site (GA2OX4=b), then the tree analyses the gene *CIB1*. Subsequently, if there is a SNP disrupting CIB1 binding site (CIB1=b), then the tree classifies the accessions as “Late” flowering phenotype without further consideration. However, if there is not a SNP disrupting *CIB1* binding site (CIB1=a), then the tree processes *FT* and the other genes in the same branch till one of the leaf nodes is reached and the decision about the phenotype classification is made. This indicates that there is a relationship between e.g. the genes *GA2OX4*, *CIB1* and *FT*. Indeed, GA2OX genes are known to regulate *FT* [34].

To answer the question of how the selected SNPs exert their effect, we examined the transcriptional relationship between the TFs whose binding sites are being affected and the target gene. This was done by co-expression analysis. We realise that this approach is not always informative because the

target gene could be controlled by many other transcription factors or its relationship is not associated with a role in the floral transition. An example of the latter case is the suppression of SOC1 by the floral organ identity proteins AP1 and AG in the flower [21]. A better approach is to study the direct transcriptional relationship by transient systems [35] or by studying the expression of target genes in mutants impaired in the TF. For most of the TFs used in this study, this information is not available yet or very incomplete and hence we relied on the co-expression relationship. By doing so, we could determine if the selected SNP results in an increase or decrease in the corresponding gene expression. To further understand the effect of the SNP on the phenotype, we examined the role of the gene, in which the SNP is found, on the control of flowering time. Subsequently, we formulated hypotheses that explain the phenotype of the accessions that carry the studied SNP.

Currently, the transcriptional relationship between TF and its target is determined based on co-expression. The inferred transcriptional relationship does not fit for, at least, the relationship between FLMβ and FT. Based on experimental evidences, FLMβ is known to physically interact with SVP to regulate repress the expression of FT [17]; while from the co-expression we can infer that FLM is an activator of FT. This is a current weak point of our approach; which can be explained by the fact that some of the transcriptome analyses have been done in mutant backgrounds. One possible route for improving the confidence of the framework proposed in this work would be by refining the approach to define the transcriptional relationship between TF and its target.

For this work, we were interested in the SNPs that potentially change the TFs binding affinity to the promoter of the flowering time genes. For that, we focused only on the SNPs that are assumed to mechanistically disrupt the regulatory region of the gene, whilst we ignored the SNPs that potentially could create a new binding site. Similarly, we also ignored the nonsynonymous SNPs located in the coding region of the TFs themselves. Relevant to note is that SNPs in the DNA binding domain of the TF leading to amino acid changes also have the potential to change the TF DNA binding affinity. In addition, in particular for the MADS TFs, which are known to require complex formation to bind to the DNA [36] and whose combinatorial interactions largely influence the DNA binding specificity [37], SNPs that overlap the protein-protein interaction domain are also important candidate as having a role in changing gene expression patterns. The dimer combination does not only influence the affinity to the target site, but also the transcriptional mode of action (activator/suppressor) may depend on the composition of the TF complex. Thus, a potential continuation of this work would be by extending the approach to include the SNPs in coding region of TFs that are involved in the flowering time control. To do so, knowledge of protein-protein interaction sites of the TFs involved [38,39] would be advantageous, and maybe as relevant as the knowledge on TF binding sites used in the current study.

In summary, the proposed approach - based on the experimental ChIP-seq data and known binding consensus motifs - was successfully applied to identify regulatory SNPs in candidate flowering genes

that have strong association with a flowering time phenotype. Previous approaches have been used already to narrow down GWAS results in order to identify specific SNPs that have an effect on gene regulation. In [10], GWAS SNPs that overlap the position of computationally identified TF binding sites are further considered as candidate regulatory SNPs. Position weight matrixes (PWMs) are used to identify the TF binding sites, and the PWMs match scores are used in the statistics to rank the importance of the SNPs in gene regulation. However, we know that only a small proportion of potential binding sites based on PWM are effectively bound by the TF [22]. In [9], SNPs overlapping the position of TF binding regions (peaks) experimentally determined by ChIP-seq studies are considered as putative regulatory SNPs. Our approach has been shown effective not only in identifying regulatory SNPs corresponding to the flowering time; but also powerful in revealing relationships between them. Additionally, we provide a simple reasoning framework that can be applied to the functional analysis of any regulatory SNP identified by our method.

Material and Methods

SNPs and flowering time data of Arabidopsis thaliana accessions

The SNP data used in this work were obtained from the 1001 Arabidopsis genomes project [40]. This involved the positions of SNPs relative to the reference genome Columbia (Col-0). Only accessions for which flowering time data is available were included in the analysis. In total, 374 accessions were used; 157 accessions from the MPICWang2013 dataset, and 217 from Salk dataset. The flowering time data were obtained from different sources: from the TAIR website [41], from the naturalvariation.org consortium [6,42,43], or from the WeigelWorld lab [44-46]. The flowering time phenotypes were binary categorized as "Early" or "Late" based on comparison against the flowering time observed for the reference Col-0 (see Table S1). We focused on SNPs spanning the gene region plus 2kb upstream of 174 known flowering time genes (listed in [11]). To determine the gene positions, the TAIR10 annotation was used.

Protein-DNA binding site data and CArG-boxes

The DNA binding sites of 11 TFs involved in plant reproduction were taken from published ChIP-seq experiments. The positions of the binding sites were taken as the ChIP-seq peaks published by the studies listed in Table 1. The genomic coordinates of all peaks were converted to be relative to the Arabidopsis genome TAIR10 using the Perl script translate_tair8.pl, as downloaded from the 1001 Arabidopsis genomes project website [40] For SVP, ChIP-seq experiment was performed in two tissues; 2 weeks old seedlings and inflorescences, representing the binding profile of SVP during vegetative and reproductive phases, respectively. In the text, these two sets of peaks are referred as SVP(vegetative) and SVP(reproductive), respectively. MADS-domain proteins are known to bind to different CArG-box sequences [22], such as the SRF-type (CC[A/T]6GG), the MEF2-type (CTA[A/T]4TAG), and other two intermediate motifs (CC[A/T]7G and C[A/T]7GG). Regular expressions were used to search for occurrence of these four types of CArG-box in both DNA strands of Arabidopsis genome TAIR10.

Predictive models based on SNPs within the flowering time genes

We used decision tree models to capture the relationship between the SNPs within the genomic region of the flowering time genes and the flowering time phenotype. To develop these models, two steps were performed: first, SNPs profiles were created, and second, these profiles were used to fit the classification tree models. A SNP profile takes the form of a matrix of binary numbers in which the columns indicate the flowering time genes and the rows indicate the accessions. Each of the binary entries in the table specifies whether or not the designated accession contains a "disruption of a regulatory region" within the genomic region of the designated gene. We identified a "disruption of a regulatory region" based on three alternative criteria: 1) if there is a SNP within a ChIP-seq peak

within the genomic region of the gene; 2) if there is a SNP within a CArG-box motif within the genomic region of the gene; or 3) if there is a SNP within a CArG-box that overlaps a ChIP-seq peak within the genomic region of the gene. A SNP profile for each of these criteria was created. Secondly, each of the three SNP profiles was used separately to fit decision tree models. In such models, the response variable corresponds to the flowering time phenotypes binary categorized as "Early" or "Late" as previously described. The model formula is Phenotype ~ Gene1 + Gene2 + … + Gene174; here *Gene I* refers to column i in the SNP profile. The models were fitted with the method rpart [47] (minsplit=30, cp=0.001), as implemented in R.

Models Assessment

Three decision tree models were defined according to each definition of "disruption of a regulatory region" of the flowering time genes, as previously described. To estimate the proportion of the phenotypic variance that can be explained by each of the models, we used them to predict the phenotype of all the accessions; then the performance of the predictions was assessed. This was performed following a leave-one-out strategy as further detailed. For each of the 374 accessions used in this work, we fitted the model using 373 accessions and predicted the phenotype for the one not used in the fitting. Then, we defined two measures to assess the predictions. The Precision was defined as *Precision=TP/*(*TP+FP*); where a true positive (*TP*) is computed when the predicted and observed phenotypes are equal to "Early", and a false positive (*FP*) is computed when the predicted phenotype is "Early" but the observed phenotype is "Late". We focus on the predictions of "Early" flowering phenotype to account for the bias in the number of accessions with "Late" flowering phenotype (66% cases have observed "Late" flowering phenotype). The Recall was defines as *Recall=TP/(TP+FN)*; where a false negative (*FN*) is computed when the predicted phenotype is equal to "Late" while the observed phenotype is equal to "Early". Then we calculated the F-score between Precision and Recall for each random set of flowering time phenotypes. For the F-score, we used the formula $F_\beta = \frac{(1+\beta^2)\times(Precision.Recall)}{\beta^2\times Precision+Recall}$; where $\beta = 1$. The frequency distribution of these was used to compute the significance of the predictions. To assess the statistical significance of the phenotype predictions, a permutation test was performed. For this test, we created 1000 sets of randomly shuffled flowering time phenotypes and compared to the predicted flowering time phenotype, such that the fraction of accessions with "Late" vs "Early" phenotypes remained the same.

Heat maps of SNP profiles

Heat maps were used to check for population structure in the data. The heat map plots were created from the obtained SNP profiles using the R method heatmap.2 (default clustering options) as implemented in the package gplots [http://cran.r-project.org/web/packages/gplots/].

Identifying SNPs with influence on the phenotype via variable importance ranking

The goal of this analysis is to use variable importance ranking to isolate a set of important SNPs that can be used to classify the accessions according to the phenotype. The ranking method varImp, as implemented in the R package caret [23], was used; then, the performance of the selection procedure was evaluated by a randomization test, as further detailed. The tree models use the formula *Phenotype ~ Gene1 + Gene2 + … + Gene174*; where *Gene i* represents a binary variable indicating whether there is a SNP disrupting a regulatory region in the genomic region of the gene *N*. Therefore, the variable importance ranks the genes according to the ability of using their associated SNPs to classify the accessions into "Early" or "Late" flowering time phenotype. The method ranks the variables according to the reduction in the loss function attributed to each variable at each split. Since there may be candidate variables that are important but are not used in a split, the top competing variables are also tabulated at each split. This means that some variables are ranked as important but are not represented by a tree split.

To evaluate the significance of the variable importance ranking analysis, we firstly created ten random sets of genes, secondly we fitted regression tree models for these sets, thirdly we performed the variable importance ranking analysis, and lastly we assessed the performance of these ten variable importance ranking. The random sets of genes were created by adding randomly 174 genes non-related to flowering time to the list of 174 flowering time genes. To assess the performance of the procedure, we calculated how many of the genes ranked as important are indeed flowering time genes; over the 10 random sets of genes, we calculated the average, minimum, maximum, 1st and 3rd quartiles of this measure to create the box plot. In addition, we used the models fitted with each of the ten random sets of genes to predict the phenotype of all the accessions; then the performances of the predictions were assessed by calculating the F-score values as previously defined. The average of F-score values obtained from the ten random sets of genes was calculated.

We next evaluated whether the variable importance values were correlated with the influence of the genes on the quality of the phenotype predictions. For that, we used the list of variables ranked as important by the model. We removed from the tree, separately, the genes ranked as the most and the less important variables; then we fitted a model without the removed gene and calculated the Precision and the Recall for the phenotype predictions. In addition, we removed from the tree all the genes selected as important by the model and we fitted a model without these genes; the Precision and the Recall for the phenotype predictions were again calculated.

Defining the role of selected SNPs in the phenotype of accessions

After having identified genes in which there is a SNP influencing the flowering time phenotype, we defined the role of that SNP. For that, we assessed if the accessions containing the selected SNP show

"Early" or "Late" flowering phenotype. For that, we used two approaches. Firstly, we analysed the tree. This was done by using the accessions as grouped by each split of the tree. For instance, for the variable in the root of the tree GA2ox4, 313 accessions have the SNP disrupting the biding site (GA2ox4=b) and 61 accessions don't carry the disruptive SNP (GA2ox4=a). From these, 69% (218 out of 313) and 52% (32 out of 61) show "Late" flowering phenotype, for GA2ox4=b and GA2ox4=a, respectively (see column 6 of Table 3). Based on these numbers, we infer that the SNP is responsible to confer "Late" flowering phenotype. The rationale behind this is that when the disruptive SNP is observed (GA2ox4=b), then the percentage of accessions with "Late" phenotype is enlarged compared to GA2ox4=b. For few cases, the gene is used in more than one split of the tree. For these, we used the split in which more accessions are observed. Secondly, for the cases in which the gene is not represented in the tree, the roles of the disruptive SNPs were determined based solely on their presence of absence among accessions with each of the phenotypes ("Early"/"Late"). Since 67% of all the used accessions show "Late" flowering phenotype, we considered that a SNP is responsible to confer a "Late" flowering phenotype if more than 67% of the accessions containing that SNP show "Late" phenotype. Similarly, a SNP is considered responsible to confer "Early" phenotype if less than 67% of the accessions containing that SNP show "Early" phenotype.

Defining regulatory relationship between the transcription factors and their targets

We used co-expression analysis to define the regulatory relationship between a transcription factor and its target. For that, we assumed that a positive pairwise co-expression correlation coefficient indicates that the transcription factor activates the expression of its target; and a negative correlation coefficient indicates that the transcription factor is a repressor of its target. To determine the co-expression correlation coefficient we used the GeneCAT webserver [48]. The correlation coefficient values between transcription factor and its target for the genes selected as important by our approach are presented in Table S1.

References

1. Rosas U, Mei Y, Xie Q, Banta JA, Zhou RW, et al. (2014) Variation in Arabidopsis flowering time associated with cis-regulatory variation in CONSTANS. Nat Commun 5: 3651.
2. Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, et al. (2010) Variation in transcription factor binding among humans. Science 328: 232-235.
3. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, et al. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. Science 342: 1235587.
4. Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, et al. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43: 956-963.
5. Brachi B, Faure N, Horton M, Flahauw E, Vazquez A, et al. (2010) Linkage and association mapping of Arabidopsis thaliana flowering time in nature. PLoS Genet 6: e1000940.
6. Atwell S, Huang YS, Vilhjalmsson BJ, Willems G, Horton M, et al. (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465: 627-631.
7. Zhang X, Cal AJ, Borevitz JO (2011) Genetic architecture of regulatory variation in Arabidopsis thaliana. Genome Res 21: 725-733.
8. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M (2012) Linking disease associations with regulatory information in the human genome. Genome Res 22: 1748-1759.
9. Bryzgalov LO, Antontseva EV, Matveeva MY, Shilov AG, Kashina EV, et al. (2013) Detection of regulatory SNPs in human genome using ChIP-seq ENCODE data. PLoS One 8: e78833.
10. Macintyre G, Bailey J, Haviv I, Kowalczyk A (2010) is-rSNP: a novel technique for in silico regulatory SNP detection. Bioinformatics 26: i524-530.
11. Fornara F, de Montaigu A, Coupland G (2010) SnapShot: Control of flowering in Arabidopsis. Cell 141: 550, 550 e551-552.
12. O'Maoileidigh DS, Graciet E, Wellmer F (2014) Gene networks controlling Arabidopsis thaliana flower development. New Phytol 201: 16-30.
13. Kaufmann K, Wellmer F, Muino JM, Ferrier T, Wuest SE, et al. (2010) Orchestration of floral initiation by APETALA1. Science 328: 85-89.
14. Yant L, Mathieu J, Dinh TT, Ott F, Lanz C, et al. (2010) Orchestration of the floral transition and floral development in Arabidopsis by the bifunctional transcription factor APETALA2. Plant Cell 22: 2156-2170.
15. Wuest SE, O'Maoileidigh DS, Rae L, Kwasniewska K, Raganelli A, et al. (2012) Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA. Proc Natl Acad Sci U S A 109: 13452-13457.
16. Deng W, Ying H, Helliwell CA, Taylor JM, Peacock WJ, et al. (2011) FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of Arabidopsis. Proc Natl Acad Sci U S A 108: 6680-6685.
17. Pose D, Verhage L, Ott F, Yant L, Mathieu J, et al. (2013) Temperature-dependent regulation of flowering by antagonistic FLM variants. Nature 503: 414-417.
18. Moyroud E, Minguet EG, Ott F, Yant L, Pose D, et al. (2011) Prediction of regulatory interactions from genome sequences using a biophysical model for the Arabidopsis LEAFY transcription factor. Plant Cell 23: 1293-1306.
19. Kaufmann K, Muino JM, Jauregui R, Airoldi CA, Smaczniak C, et al. (2009) Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower. PLoS Biol 7: e1000090.
20. Gregis V, Andres F, Sessa A, Guerra RF, Simonini S, et al. (2013) Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in Arabidopsis. Genome Biol 14: R56.
21. Immink RG, Pose D, Ferrario S, Ott F, Kaufmann K, et al. (2012) Characterization of SOC1's central role in flowering by the identification of its upstream and downstream regulators. Plant Physiol 160: 433-449.
22. de Folter S, Angenent GC (2006) trans meets cis in MADS science. Trends Plant Sci 11: 224-231.

23. Kuhn M (2008) Building Predictive Models in R Using the caret Package. Journal of Statistical Software 28.
24. Kim J, Kim Y, Yeom M, Kim JH, Nam HG (2008) FIONA1 is essential for regulating period length in the Arabidopsis circadian clock. Plant Cell 20: 307-319.
25. Corbesier L, Vincent C, Jang S, Fornara F, Fan Q, et al. (2007) FT protein movement contributes to long-distance signaling in floral induction of Arabidopsis. Science 316: 1030-1033.
26. Alice Pajoro SB, Evangelia Dougali, Felipe Leal Valentim, Marta Adelina Mendes, Aimone Porri, George Coupland, Yves Van de Peer, Aalt D.J. van Dijk, Lucia Colombo, Brendan Davies and Gerco C. Angenent (2014) The (r)evolution of gene regulatory networks controlling *Arabidopsis* plant reproduction, a two decades history. Journal of Experimental Botany Accepted.
27. Rieu I, Ruiz-Rivero O, Fernandez-Garcia N, Griffiths J, Powers SJ, et al. (2008) The gibberellin biosynthetic genes AtGA20ox1 and AtGA20ox2 act, partially redundantly, to promote growth and development throughout the Arabidopsis life cycle. Plant J 53: 488-504.
28. Fornara F, Panigrahi KC, Gissot L, Sauerbrunn N, Ruhl M, et al. (2009) Arabidopsis DOF transcription factors act redundantly to reduce CONSTANS expression and are essential for a photoperiodic flowering response. Dev Cell 17: 75-86.
29. Hudson TJ (2003) Wanted: regulatory SNPs. Nat Genet 33: 439-440.
30. Wasserman WW, Sandelin A (2004) Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 5: 276-287.
31. Muino JM, Smaczniak C, Angenent GC, Kaufmann K, van Dijk AD (2014) Structural determinants of DNA recognition by plant MADS-domain transcription factors. Nucleic Acids Res 42: 2138-2146.
32. Zou H, T. H (2005) Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society B 67: 301–320.
33. Kursa MB (2014) Robustness of Random Forest-based gene selection methods. BMC Bioinformatics 15: 8.
34. Song YH, Ito S, Imaizumi T (2013) Flowering time regulation: photoperiod- and temperature-sensing in leaves. Trends Plant Sci 18: 575-583.
35. Wehner N, Hartmann L, Ehlert A, Bottner S, Onate-Sanchez L, et al. (2011) High-throughput protoplast transactivation (PTA) system for the analysis of Arabidopsis transcription factor function. Plant J 68: 560-569.
36. Schwarz-Sommer Z, Huijser P, Nacken W, Saedler H, Sommer H (1990) Genetic Control of Flower Development by Homeotic Genes in Antirrhinum majus. Science 250: 931-936.
37. Smaczniak C, Immink RG, Muino JM, Blanvillain R, Busscher M, et al. (2012) Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. Proc Natl Acad Sci U S A 109: 1560-1565.
38. van Dijk AD, Morabito G, Fiers M, van Ham RC, Angenent GC, et al. (2010) Sequence motifs in MADS transcription factors responsible for specificity and diversification of protein-protein interaction. PLoS Comput Biol 6: e1001017.
39. Leal Valentim F, Neven F, Boyen P, van Dijk AD (2012) Interactome-wide prediction of protein-protein binding sites reveals effects of protein sequence variation in Arabidopsis thaliana. PLoS One 7: e47022.
40. Weigel D, Mott R (2009) The 1001 genomes project for Arabidopsis thaliana. Genome Biol 10: 107.
41. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, et al. (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res 40: D1202-1210.
42. Li Y, Huang Y, Bergelson J, Nordborg M, Borevitz JO (2010) Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana. Proc Natl Acad Sci U S A 107: 21199-21204.

43. Platt A, Horton M, Huang YS, Li Y, Anastasio AE, et al. (2010) The scale of population structure in Arabidopsis thaliana. PLoS Genet 6: e1000843.
44. Balasubramanian S, Sureshkumar S, Lempe J, Weigel D (2006) Potent induction of Arabidopsis thaliana flowering by elevated growth temperature. PLoS Genet 2: e106.
45. Lempe J, Balasubramanian S, Sureshkumar S, Singh A, Schmid M, et al. (2005) Diversity of flowering responses in wild Arabidopsis thaliana strains. PLoS Genet 1: 109-118.
46. Werner JD, Borevitz JO, Uhlenhaut NH, Ecker JR, Chory J, et al. (2005) FRIGIDA-independent variation in flowering time of natural Arabidopsis thaliana accessions. Genetics 170: 1197-1207.
47. Terry Therneau, Beth Atkinson, Ripley B (2014) Recursive partitioning for classification, regression and survival trees. An implementation of most of the functionality of the 1984 book by Breiman, Friedman, Olshen and Stone.
48. Mutwil M, Obro J, Willats WG, Persson S (2008) GeneCAT--novel webtools that combine BLAST and co-expression analyses. Nucleic Acids Res 36: W320-326.
49. El-Din El-Assal S, Alonso-Blanco C, Peeters AJ, Raz V, Koornneef M (2001) A QTL for flowering time in Arabidopsis reveals a novel allele of CRY2. Nat Genet 29: 435-440.
50. Aukerman MJ, Sakai H (2003) Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. Plant Cell 15: 2730-2741.
51. Liu H, Yu X, Li K, Klejnot J, Yang H, et al. (2008) Photoexcited CRY2 interacts with CIB1 to regulate transcription and floral initiation in Arabidopsis. Science 322: 1535-1539.
52. Aubert D, Chen L, Moon YH, Martin D, Castle LA, et al. (2001) EMF1, a novel protein involved in the control of shoot architecture and flowering in Arabidopsis. Plant Cell 13: 1865-1875.
53. Laubinger S, Marchal V, Le Gourrierec J, Wenkel S, Adrian J, et al. (2006) Arabidopsis SPA proteins regulate photoperiodic flowering and interact with the floral inducer CONSTANS to regulate its stability. Development 133: 3213-3222.
54. Koornneef M, van der Veen JH (1980) Induction and analysis of gibberellin sensitive mutants in Arabidopsis thaliana (L.) heynh. Theor Appl Genet 58: 257-263.
55. Streitner C, Danisman S, Wehrle F, Schoning JC, Alfano JR, et al. (2008) The small glycine-rich RNA binding protein AtGRP7 promotes floral transition in Arabidopsis thaliana. Plant J 56: 239-250.
56. Proveniers M, Rutjens B, Brand M, Smeekens S (2007) The Arabidopsis TALE homeobox gene ATH1 controls floral competency through positive regulation of FLC. Plant J 52: 899-913.

Table 3 - Overview of mutations with effect on the regulatory network of flowering time control.

Gene[1]		VarImp[2]	CArG-box[3]		ChIP-seq TF[4]	Regulatory relationship[5]	Tree analysis[6]	Inferred SNP role[7]	Evidence-driven inferred SNP role[8]	Mutant phenotype[9]	Role of gene in the induction of floral transition[10]
AT2G21070	FIO1	7.95	GAATATAACC	Chr2:9042713-9042723	SVP(veg.)	SVP:repressor	a=50% Late p=66% Late	Late	Late	knock-out mutant early flowering in LD and SD [24]	Floral Repressor
AT1G04400	CRY2	5.15	GGATTAAATC	Chr1:11881103-1188113	AP1	AP1:activator	a=58% Late b=85% Late	Late	Late	Mutant late flowering in LDs [49]	Floral Inducer
AT2G28550	TOE1	4.40	CCTTATTAGG	Chr2:12225911-12225921	AP1	AP1:repressor	---	Early*	unclear (AP1 acts after floral transition)	knock-out mutant early flowering in LDs [50]	Floral Repressor
AT1G65480	FT	3.58	CCTTTTTTGGG	Chr1:24331801-24331811	AP1, FLC, FLM	AP1:activator FLC:repressor FLM:activator$^{\infty}$	a=67% Late p=81% Late	Late	ambiguous (disruption of activators and repressors)	knock-out mutant late flowering in LDs [25]	Floral Inducer
AT2G34555	GA2OX3	3.45	GGAAAAAAACC	Chr2:14557657-14557668	SVP(rep.)	SVP:repressor	a=62% Late p=73% Late	Late	Late	single knock-out mutant has no flowering phenotype; double, triple and quintuple *ga2ox* mutants flower early under short days [27]	Floral Repressor
AT1G47990	GA2OX4	3.01	GAAAAAAACC	Chr1:17699775-17699785	AG, AP1, PI	AG:repressor AP1:activator PI:activator	a=69% Late p=52% Late	Early	unclear (AG, AP1 and PI act after floral transition)	single knock-out mutant has no flowering phenotype; double, triple and quintuple *ga2ox* mutants flower early under short days [27]	Floral Repressor
AT2G18915	LKP2	2.03	CCAAAAATTG	Chr2:8198467-8198477	SVP(veg.)	SVP:activator	a=63% Late p=40% Late	Early	Early	*fkf1 ztl lkp2* triple mutants are late flowering in LDs; LKP2 overexpressor late flowering in LD [28]	Floral Inducer
AT4G22950	AGL19*	1.76	CCAAATAAGG	Chr4:12026611-12026621	AG, AP1, AP3, PI	AP1:repressor AP3:repressor	---	Early*	undetermined	---	---
AT4G34530	CIB1	1.71	CTATTTATAG	Chr4:6499569-16499579	SVP(rep.)	---	a=48% Late p=70% Late	Late	undetermined	*cib1 cib5* double mutant shows delayed flowering; CIB1 overexpression causes early flowering [51]	Floral Inducer
AT3G24440	VIL1	1.23	CAAAAAAAGG	Chr3:8878520-8878530	FLM	FLM:activator	a=59% Late p=79% Late	Late	undetermined	---	---
AT5G11530	EMF1	1.22	CCTAAAATAG	Chr5:3694271-3694281	AP1, AP3, PI	AP1:activator AP3:activator PI:activator	a=59% Late p=75% Late	Late	Early	knock-out mutants are extremely early flowering, flowering as seedlings [52]	Floral Repressor
AT3G15354	SPA3	0.94	CCTTTTTTTG	Chr3:5173179-5173189	AP1,SEP3	---	---	---	undetermined	*spa1 spa3 spa4* triple mutants flower early in LDs and SDs [53]	Floral Repressor
AT4G02780	CPS1	0.68	GTAAATAACC	Chr4:1241666-1241676	SVP(veg.)	SVP:repressor	---	Early*	Early	knock-out mutant late flowering in long days[54]	Floral Inducer
AT2G21660	ATGRP7	0.66	CATAATTTGG	Chr2:9267122-9267132	SVP(veg.)	---	---	---	undetermined	mutant late flowering [55]	Floral Inducer
AT4G32980	ATH1*	0.63	CCTAAAAAAG	Chr4:15919701-15919711	SVP(veg.)	SVP:activator	---	Early*	Early	knock-out mutant early flowering [56]	Floral Repressor
AT1G14920	GAI	0.53	GGATAATTTC	Chr1:5150346-5150356	SVP(veg.)	SVP:activator	---	---	---	Late flowering in short days[55]	Floral Inducer

* Since these genes are not represented in the tree, the roles of the disruptive SNPs were determined based solely on their presence or absence in accessions with one of the phenotypes ("Early"/"Late"); $^{\infty}$Relationship based on co-expression. Experimental evidences indicate that this relationship depends on the FLM splicing variant. [1]Gene ID and gene ID. [2]Variable importance ranking value. [3]CArG-box sequence and CArG-box genomic position.The bold nucleotides indicate the position of the SNP [4]TFs whose binding sites are being disrupted by the SNP. [5]Regulatory relationship between the TF and target gene, as inferred by co-expression analysis. [6]*p* and *a* stand for, respective, presence and absence of the disruptive SNP. [7]The SNP role is inferred by comparing the percentages of accessions that have "Late" flowering phenotype when gene has the SNP (=b) or when it does not have (=a). [8]Role of the SNP as inferred by experimental evidences (co-expression, flowering phenotype). [9]Information about mutant phenotype. [10]Role of the target gene on floral transition.

Supporting Information

Figure S1: Graphical representation of the SNPs over the genic region of the gene SPA3. The gene structure is represented by the bar on top of the figure, where black rectangles delimitate the region of the exons, and grey rectangles delimitate the 2kb upstream the gene region. The positions of ChIP-seq peaks of the 11 TFs involve in plant development are indicated by the red columns. The positions of the CArG-boxes are shown by the blue columns. The SNPs of different accessions are represented by the black or grey columns, representing accessions that show "Early" or "Late" flowering time respectively. Only the 25 accessions which have a SNP overlapping a CArG-box that is located within a ChIP-seq peak are represented.

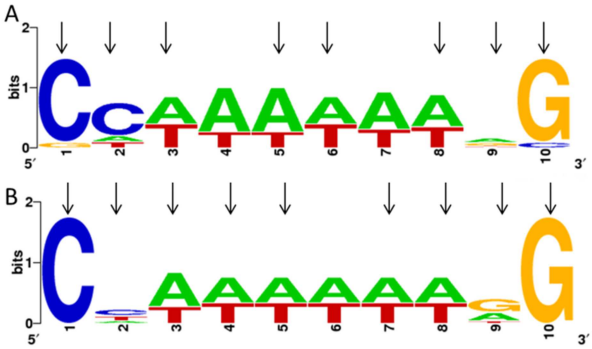


Figure S2: Graphical representation of the CArG-boxes that are located within ChIP-seq peaks and in which there is a SNP. (A) Logo constructed with the CArG-boxes of the 15 selected genes. **(B)** Logo constructed with the CArG-boxes of the 55 genes that have a disruption in the CArG-box located within the ChIP-seq peak but were not selected as having a strong association with the flowering time phenotype. The positions of the disruptive SNPs are indicated by the arrows.

Table S1: Co-expression correlation coefficient values between transcription factor and its target for the genes selected as important by our approach.

Target	Co-expression(TF,Target)
FIO1	SVP: -0.19
FT	AP1: 0.35 FLC: -0.021 FLM: 0.22
GA2OX4	AG: -0.020 AP1: 0.019 PI: 0.035
GA2OX3	SVP: -0.25
TOE1	AP1: -0.40
CRY2	AP1: 0.12
SPA3	AP1:not found SEP3:not found
VRN5	FLM: -0.23
CPS1	SVP: -0.13
AGL19	AG: not found AP1: -0.26 AP3: -0.34 PI: not found
EMF1	AP1: 0.19 AP3: 0.33 PI: 0.24
CIB1	SVP: not found
ATH1	SVP: 0.31
GRP7	SVP: not found
LKP2	SVP: -0.033
VIL1	FLM: 0.175
GAI	SVP: 0.558

Chapter 4

Interactome-Wide Prediction of Protein-Protein Binding Sites Reveals Effects of Protein Sequence Variation in Arabidopsis thaliana

Felipe Leal Valentim, Frank Neven, Peter Boyen and Aalt D. J. van Dijk

A modified version is published in *Plos One* (2012) 7(10): e47022

Abstract

The specificity of protein-protein interactions is encoded in those parts of the sequence that compose the binding interface. Therefore, understanding how changes in protein sequence influence interaction specificity, and possibly the phenotype, requires knowing the location of binding sites in those sequences. However, large-scale detection of protein interfaces remains a challenge. Here, we present a sequence- and interactome-based approach to mine interaction motifs from the recently published *Arabidopsis thaliana* interactome. The resultant proteome-wide predictions are available via www.ab.wur.nl/sliderbio and set the stage for further investigations of protein-protein binding sites. To assess our method, we first show that, by using *a priori* information calculated from protein sequences, such as evolutionary conservation and residue surface accessibility, we improve the performance of interface prediction compared to using only interactome data. Next, we present evidence for the functional importance of the predicted sites, which are under stronger selective pressure than the rest of protein sequence. We also observe a tendency for compensatory mutations in the binding sites of interacting proteins. Subsequently, we interrogated the interactome data to formulate testable hypotheses for the molecular mechanisms underlying effects of protein sequence mutations. Examples include proteins relevant for various developmental processes. Finally, we observed, by analysing pairs of paralogs, a correlation between functional divergence and sequence divergence in interaction sites. This analysis suggests that large-scale prediction of binding sites can cast light on evolutionary processes that shape protein-protein interaction networks.

Introduction

Genotype-to-phenotype relationships are mediated via molecular networks, including protein-protein interaction networks. Hence, understanding how phenotypes are influenced by sequence changes requires understanding how the specificity of protein interactions is encoded in protein sequences. Identifying which sites are involved in the interactions is a necessary step towards studying the underlying molecular mechanisms and the evolutionary processes influencing protein interaction networks. However, accurate automatic detection of protein binding sites remains a challenge when aiming at large-scale identification.

Those interaction sites composing the protein interface are directly identifiable given a 3D structure of a complex [1]; when only the unbound protein structure is known, predictions based on structural and physicochemical properties [2], [3], [4] are typically used. Although very relevant, protein structure determination is not able to cover the large number of interactions identified by interactome projects [5]. In particular for plants, including the model plant species *Arabidopsis thaliana*, there is a gap between the amount of protein-protein interactions experimentally unravelled and the amount of structural information available in the Protein Data

Bank [6]. This gap highlights the need for sequence-based approaches for large-scale predictions of interfaces.

Recently, the Arabidopsis Interactome map has been released, describing about 6,200 highly reliable interactions between about 2,700 proteins [7]. Due to the high rate of gene duplication in the Arabidopsis genome [8], [9], it is particularly interesting to investigate the relationship between protein interaction specificity and sequence diversity in Arabidopsis proteins: after duplication, interaction specificity can diverge causing non-, sub- or neo-functionalization [10]. However, the relationship between interaction specificity and sequence similarity is far from trivial. For example, when analysing pairs of yeast duplicated genes [11] changes in interaction specificity were not correlated with sequence divergence, when this divergence was calculated over the whole length of the protein sequence. Locating the protein-protein binding sites of several duplicated genes may create new routes for this type of investigation, since it would enable to evaluate selective pressure specifically in functional parts of the sequence.

In contrast to protein structures, in which an interaction site is seen as a continuous stretch of amino acids in space, protein sequences show an interface as scattered short sub-sequences. It has been suggested that proteins with common interaction partners also share common functional features [12], such as the short sequences composing the interface. Still, these shared motifs are difficult to discover, perhaps due to their short length. It has also been shown that evolutionary conservation may be useful in predicting functional motifs in the protein surface [13], [14], but for discriminating protein-protein interfaces from other functional sites, e.g. small ligand binding sites and catalytic sites, its use as a stand-alone predictor is questionable [15]. In this work, we evaluate the performance of an interactome-based interaction site predictor when information encoded in the protein sequences is included in its calculation.

We previously developed a method that uses protein-protein interaction networks to find sequence motifs shared by proteins with common interaction partners [16]. This method outperformed existing correlated motif mining algorithms and was able to find biologically meaningful motifs from large protein-protein interaction networks. Here, we present a version of the method modified to account also for the evolutionary conservation of homologous sequences. In addition, the method proposed here restricts the motif search to sequence regions that are likely to be exposed in the protein surface. This new sequence- and interactome-based method predicts motifs that are not only shared by proteins with common interaction partners, but also conserved across sequences of orthologs in closely related species and likely to be exposed in the protein surface.

We start by assessing the performance of our new method. By comparing our predictions against available structural information, we show that the modifications in the method improve its performance. In addition, the assessment provides a basis for determining a set of default parameters for the algorithm. Next, we obtain large-scale predictions of protein interaction sites from the complete Arabidopsis interactome data. We use single nucleotide polymorphism data to obtain evidence that the predicted binding sites are functionally relevant. Subsequently, we analyse available data describing the effect of amino acid mutagenesis to show that our predictions can be interrogated to obtain insight into previously unknown molecular mechanisms underlying the effect of specific mutations. Finally, we analyse the sequences of paralogous pairs to set the stage for further investigations of the molecular mechanisms behind the link between sequence diversity and functional divergence in Arabidopsis proteins.

Results and Discussion

SLIDERBio algorithm

We recently developed SLIDER, a method that uses a protein interaction network to locate binding sites in the sequence of interacting proteins [16]. To predict binding sites for the proteins in the recently generated Arabidopsis interactome [7], we modified this algorithm to enable it to take various types of biological knowledge into account. Here, we give a brief overview of the method, focusing on the modifications that lead to a novel algorithm. Our method follows the assumption that interfaces can be represented by short sequence motifs (Figure 1). To predict such motifs, the algorithm mines a set of sequences of interacting proteins aiming to find motif pairs overrepresented in pairs of interacting proteins. This mining results in a set of motif pairs that are predicted to be located in protein-protein interfaces. For this work, we extended the original SLIDER algorithm by implementing a different approach to define the presence of a motif in a sequence, and by adding additional filtering steps based on the evolutionary conservation and surface accessibility predicted from the protein sequences. This new, improved version is hereafter named SLIDERBio and is available for download at www.ab.wur.nl/sliderbio.

For computational details of the SLIDER method, the reader is referred to [16]. In summary, the algorithm makes use of an objective function that quantifies the overrepresentation of a motif pair based on its presence in pairs of interacting proteins. To start, it randomly selects a short motif from protein sequences. To optimize the objective function, the algorithm heuristically "slides" the position of the selected motif. This method has been shown to outperform existing methods for mining binding motifs from interaction networks [16].

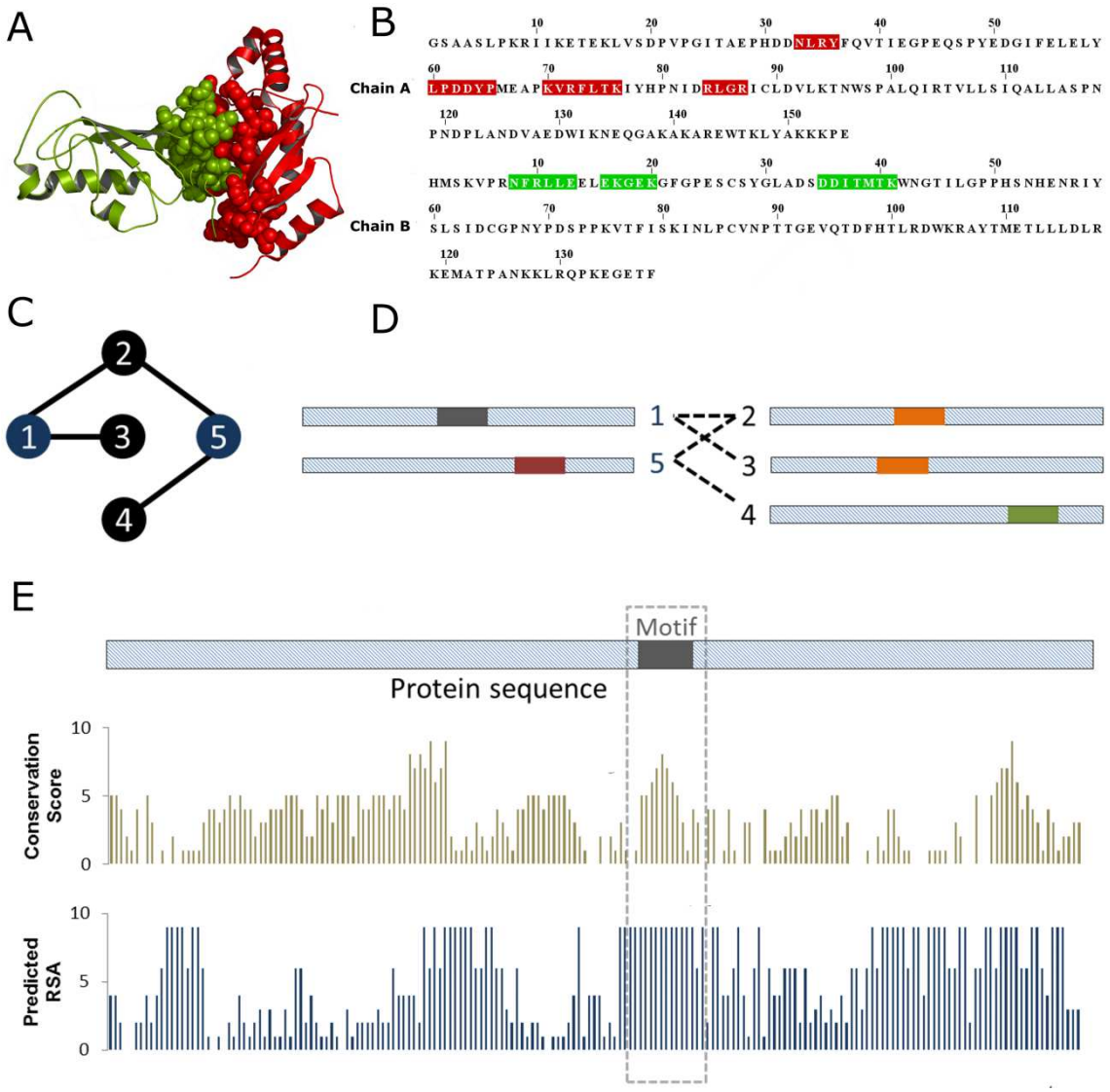


Figure 1: SLIDERBio strategy to predict protein-protein binding sites. (A–B) SLIDERBio follows the assumption that interfaces can be represented by short sequence motifs: (A) Interaction sites (spacefill) are continuous patches of amino acid residues in the 3D structure of a protein, while in a protein sequence (B) the interface is composed of scattered short motifs (regions highlighted in red and green). In (A–B), protein structure and sequence of the Mms2/Ubc13 heterodimer (PDB id 1jat) are used as illustration. (C–D) SLIDERBio predicts interaction sites by finding motif pairs that are overrepresented in pairs of interacting proteins in an interaction network. (C) illustrates a protein-protein interaction network in which the proteins are represented by nodes and the interactions represented by connecting edges; (D) illustrates the protein sequences and their short motifs (regions highlighted in colored bars; same colors represents similar motifs). In this example, the motif pair [grey-orange] is overrepresented compared to the motif pair [red-green]. To calculate the degree of overrepresentation of a motif, the method verifies in how many sequences of interacting proteins a certain motif is found. Originally, SLIDER considered a motif present in a sequence if a perfect match was found between motif sequence and a region in the protein sequence. In contrast, SLIDERBio makes use of a substitution matrix to calculate the similarity between the motif and the sequence. If the degree of similarity between a motif

and a sequence is greater than a threshold, SLIDERBio considers that the sequence contains the motif. In addition, SLIDERBio verifies whether the conservation score and the surface accessibility score of the motifs are greater than pre-defined thresholds. These three thresholds are based on the average value per residue over the length of the motif (E).

One critical step in the algorithm consists of verifying whether a short motif is present in a protein sequence. Originally, SLIDER considered that a protein contained a motif if a perfect match was found between motif sequence and a region in the protein sequence. In contrast, the SLIDERBio algorithm makes use of the BLOSUM62 [17] substitution matrix to derive a value that reflects the degree of similarity between the motif and the sequence (see Materials and Methods). In other words, the original SLIDER scanned the protein sequences searching for a perfect match for a motif sequence, while the SLIDERBio algorithm searches for a "close" match. This degree of similarity calculated using the substitution matrix reflects "how close" the match is. Only if the degree of similarity between a motif and a sequence is greater than a threshold, then SLIDERBio considers that the sequence contains the motif.

Additionally, to select only those overrepresented motifs that are likely to be located in the interaction interface, filtering steps based on pre-calculated biological information were implemented. SLIDER considered that a protein contained all the motifs that satisfy the sequence match criteria. For SLIDERBio, the region from the protein sequence that matches the motif has to satisfy two extra conditions: (i) it has to show evolutionary conservation greater than a conservation threshold, and (ii) it has to have predicted surface accessibility greater than an accessibility threshold (Figure 1D). These requirements are based on the fact that interface residues should be located at the surface of a protein (i.e. have high enough accessibility) and that compared to surface residues that are not involved in functions such as protein binding, they are expected to have higher conservation. To implement these filtering steps, the method compares the averages of predicted residue conservation and residue accessibility score calculated over the length of the overrepresented motifs to their thresholds. The strategies to calculate the conservation score and residue surface accessibility are discussed in the Materials and Methods section. Briefly, conservation is assessed using an entropy based score, and residue surface accessibility is predicted using a neural network approach. Values obtained from both approaches are rescaled in the range 0 to 9, and SLIDERBio applies a threshold on those rescaled values. The analysis presented in the section Assessment of SLIDERBio predictions allows determining the best set of threshold values.

Before the modifications, SLIDER required as input only protein sequences and protein-protein interaction data. The SLIDERBio algorithm now additionally requires the conservation score and the predicted surface accessibility for all proteins. In addition, SLIDERBio requires the user to set values for parameters that determine the thresholds of degree of similarity, conservation

and residue solvent accessibility. The performance of various parameter settings was analysed by comparing our sequence-based SLIDERBio predictions with available protein structure data. This analysis allowed to assess the significance of the inclusion of the biological information in SLIDERBio and, furthermore, to obtain a default set of parameters. Next, we predicted protein interaction motifs for the Arabidopsis interactome and investigated the predicted interaction sites, in particular aiming at applying these towards understanding the effect of sequence variation.

Assessment of SLIDERBio predictions

We analysed SLIDERBio predictions aiming (i) to assess the performance of the algorithm towards large scale predictions of protein binding motifs; (ii) to evaluate the significance of the implemented modifications and; (iii) to obtain a set of default values for the parameters. For these investigations, we used a subset of protein-protein interactions such that for the proteins involved, their sequences could be mapped to available structures of protein complexes; hence the interface residues could directly be identified for assessment of our predictions. Hereafter, these subsets are referred to as "structurally mapped datasets". Although we focus our application on *Arabidopsis thaliana*, for this assessment, given the small number of Arabidopsis proteins with structural mapping, we also used human and yeast protein-protein interaction data (see Figure S1; Tables S1 andS2). We tested SLIDERBio on the structurally mapped datasets of the three species using 180 different parameter settings. To analyse the results, we defined two measures that quantify the quality of the predictions: "Accuracy of predicted motifs" and "Coverage of protein-protein interfaces" (see Materials and Methods). Both measures were combined into an F-score (harmonic mean of Accuracy and Coverage) as overall performance measure.

Firstly, we observed that for most of the parameter settings, SLIDERBio obtains better results than the previous SLIDER, in terms of both Accuracy and Coverage (Figure 2, A–C). Note that our previous analysis of SLIDER already showed that it obtained improved performance compared to existing correlated motif mining algorithms. Depending on the parameter values, SLIDERBio could predict motifs with Coverage of protein-protein interfaces up to 42%, 22% and 42%, respectively for the human, yeast and Arabidopsis subsets. Likewise, the values of Accuracy of predicted motifs were up to 58%, 96% and 100%. We focus the subsequent analyses based on the F-scores, which give a compromise between 'Accuracy of predicted motifs' and 'Coverage of protein-protein interfaces'.

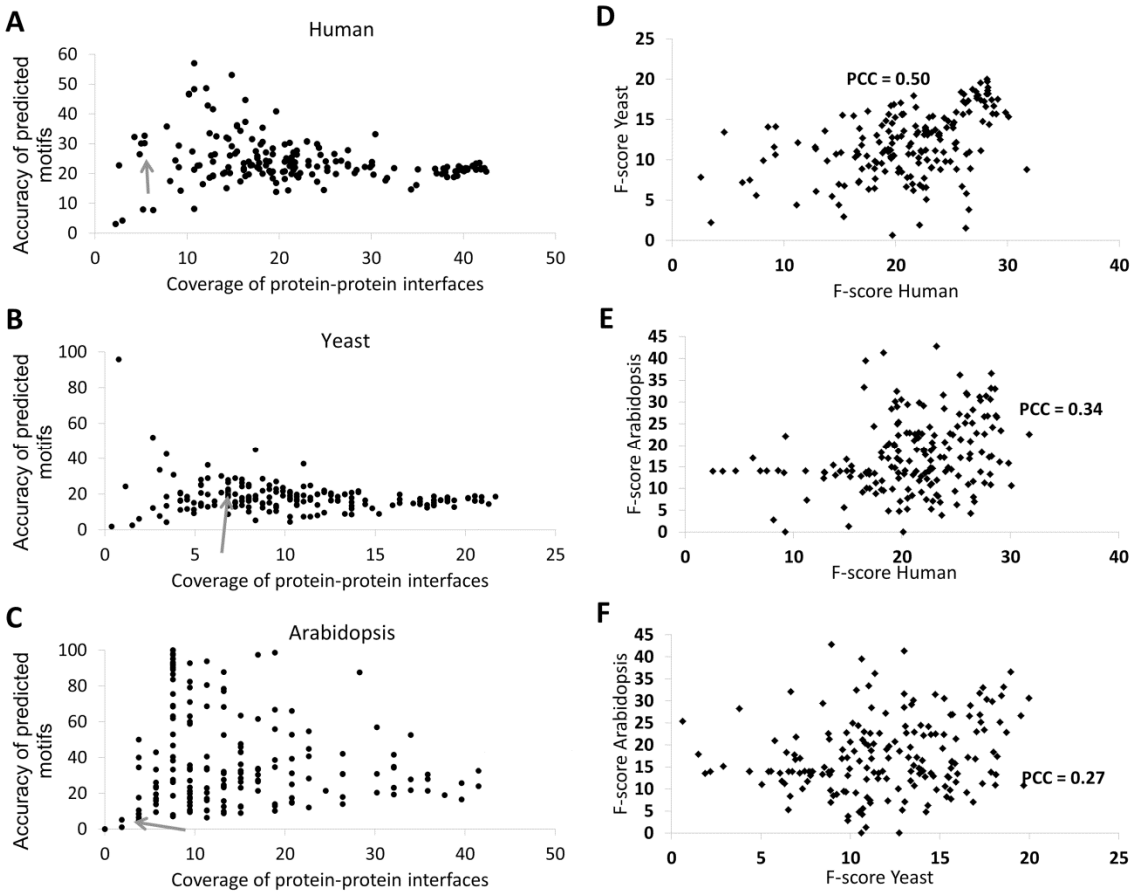


Figure 2: Overall performance of the SLIDERBio algorithm in different datasets. (A–C) Coverage of protein-protein interfaces and Accuracy of predicted motifs. Each dot represents the result of SLIDERBio using one of the 180 tested sets of parameters, for (A) human, (B) yeast and (C) Arabidopsis structurally mapped subsets. The grey arrows indicate the dot corresponding to the result of the previous SLIDER algorithm. (D–F), Correlation of the performance for each of the SLIDERBio parameter settings is compared among datasets of different species: (D) human vs. yeast; (E) human vs. Arabidopsis; and (F) yeast vs. Arabidopsis. Pearson Correlation Coefficient (PCC) is indicated.

Secondly, scatter diagrams and Pearson's correlation coefficients (PCC) were used to determine whether F-scores obtained for the same parameter settings are correlated among the three structurally mapped datasets. A strong correlation implies here that the same set of parameters would give results with similar quality in different datasets. A good correlation is particularly important, because we based our assessment on structurally mapped datasets of three species in order to determine the best parameter setting for further predictions on the complete Arabidopsis interactome data. When comparing the F-scores obtained for the same parameters but networks from different species (comparison shown in Figure S1 and S2), we found significant positive correlation: PCC = 0.50, PCC = 0.34 and PCC = 0.27, for correlation of results from human/yeast, human/Arabidopsis and yeast/Arabidopsis, respectively (Figure 2, D–F). From the data inFigure S1 and S2, it is apparent that there is more similarity between the

degree distribution of the human and yeast structurally mapped datasets and the complete Arabidopsis interactome than between the Arabidopsis structurally mapped dataset and the complete Arabidopsis interactome. Hence, a reason for the observed smallest correlation between the results in Arabidopsis with those in yeast and human might be that the topology of the structurally mapped Arabidopsis dataset differs most from the other two. In addition, it might also be because of the fact that the number of structurally mapped proteins in the Arabidopsis dataset is much smaller than those of the other species, leading to a larger variation in apparent performance. Overall, the good correlation between the F-scores indicates that parameters that give good results for all three structurally mapped datasets, would also give good results for the complete Arabidopsis interactome.

Thirdly, boxplots were used to group the F-score results according to the used threshold values, thus allowing assessment of the significance of each modification isolated from the effect of the other modifications. The most striking result from this assessment is that, in all the three species, the inclusion of the residue surface accessibility information significantly improved the quality of the results (p-value <0.01, paired t-test; Figure S3). Moreover, the highest value of the surface accessibility threshold (value 7) resulted in the highest F-scores, independently of the values that were used for the other two thresholds.

Lastly, we conducted randomization tests to quantify the significance of our results regarding the F-scores, and in addition, to determine the best set of parameters. To obtain$p$-values, we compared the SLIDERBio results against 1,000 sets of randomly generated motif pairs (see Materials and Methods). We selected parameter settings for further consideration using a significance level threshold of p-value <0.05 (Figure S4). Note that *a priori* we do not necessarily expect a lot of parameter settings to show significant results, because several parameter combinations will combine biological information in a non-optimal way: e.g. when the threshold for conservation is high and the threshold for accessibility is low, we expect to predict a lot of buried conserved residues instead of interface residues. Although eight parameter settings showed p-values less than 0.05 simultaneously for the human and yeast predictions, only one occurred simultaneously for all the three species. Hence, we selected this combination of parameters [Degree of similarity = 0.6; Conservation = 6; Surface accessibility = 7] as the setting to run SLIDERBio for predictions on the full Arabidopsis interactome. These values for the parameters mean that for a motif to occur in a sequence it has to have an average similarity of at least 60%, and that the residue conservation score and residue surface accessibility score have on average values greater than 6 and 7, respectively.

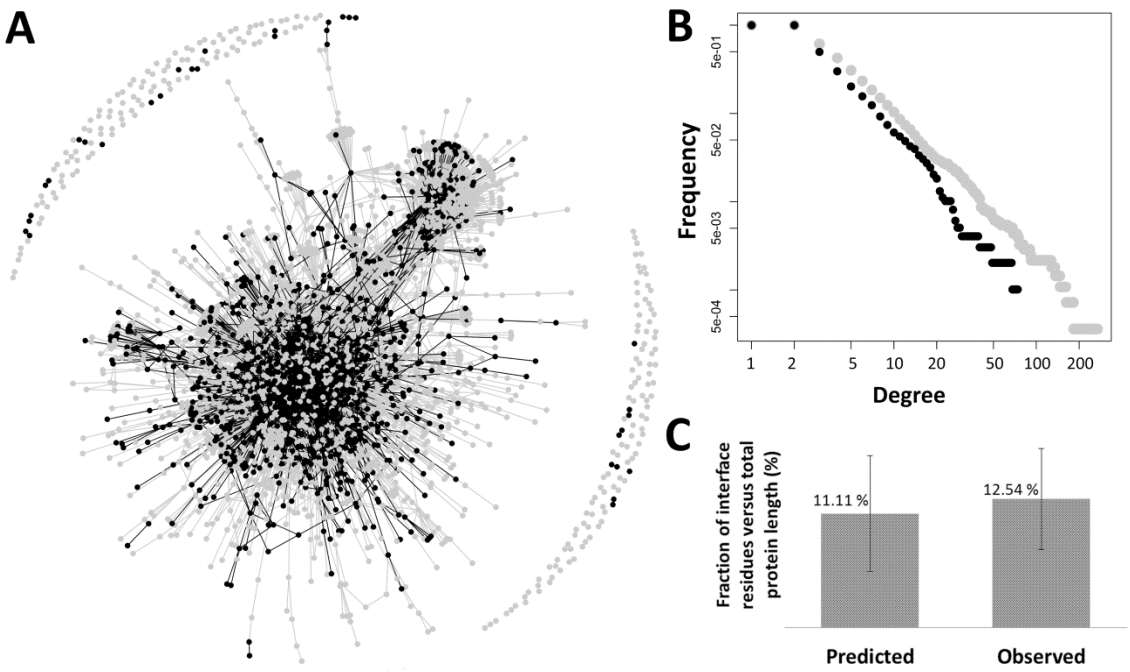


Figure 3: Overall description of the predicted binding sites in the Arabidopsis interactome. (A) Network representation of the Arabidopsis interactome and predicted interaction sites. The vertices and edges in black represent, respectively, the 985 proteins and the 1498 interactions to which predicted motifs are mapped. (B) Degree distributions from the complete protein-protein interaction dataset (grey) and from the subset with only proteins and interactions that have a predicted motif (black). A and B suggest that our method is not biased to predict motifs that can be mapped only to proteins with high degree (*i.e.* number of interactions); moreover, the proteins with predicted motifs are distributed in different positions in the network. (C) Percentage of residues in the interfaces, either in the predicted interfaces or those observed in the structurally mapped dataset. Standard deviation is indicated.

Turning now to the complete Arabidopsis interactome data, our method predicted protein-protein binding motifs that could be mapped (See Materials and Methods) to 1498 (24%) of the interactions among 985 (36%) proteins distributed over the entire network (Figure 3A). Comparison of the degree distribution from the complete dataset against the degree distribution from the subset of proteins with a predicted binding site suggests that the method is not biased to identify motifs only in those proteins with high number of interactions (Figure 3B). Moreover, the motifs mapped onto the protein sequences cover on average 11% of the total protein length, which is a reasonable number given that the equivalent percentage based on protein complexes structures comprising the Arabidopsis structurally mapped dataset is 12%

(Figure 3C). For each protein, the resulting predicted sites are given in Table S3; these are also available via www.ab.wur.nl/sliderbio. In addition, for each interaction listed in the interactome, the motif pair(s) predicted to be responsible for the interaction is given. This set of predictions, which is comprised by motifs that are overrepresented in pairs of interacting proteins, conserved across species and predicted to be located in the surface of the protein structure, was used for further analysis.

Protein-protein binding sites variability and intermolecular coevolution

Conserved residues exposed in the surface of the protein are likely be involved in its biological activity. To obtain an indication of the functional relevance of the predicted binding sites, we used single nucleotide polymorphism (SNP) data (i.e. conservation within *Arabidopsis thaliana*). If our predicted interaction sites are indeed functionally important, one would expect less variability in their positions compared to the rest of the protein sequence. To test this hypothesis, we calculated the percentage of predicted interface residues in which a non-synonymous SNP (nsSNP) is found (1.6%); this is significantly lower than the percentage of all protein residues in which a nsSNP is found (2.2%; *p*-value<0.001; see Materials and Methods). As a control, we tested that a similar signal was not obtained when using synonymous SNPs (data not shown).

Those nsSNP that are found in regions of predicted binding sites are potentially interesting because, by changing protein interaction specificity, they might be responsible for conferring variability to different individuals of a species. However, considering evidence that most interactions are conserved within species [18], one would expect that when an interaction site is mutated, there might be a tendency to have compensating mutations in the interaction partners. Such scenario is consistent with the intermolecular co-evolution model [19]. In our case, it leads to the hypothesis that proteins in which an nsSNP is found overlapping a predicted binding site would be expected to have an increased tendency to interact with other proteins in which an nsSNP is also found in a binding site. To test this hypothesis we counted the number of interactions between proteins in which a nsSNP overlaps a binding site, from which we found a number significantly greater than what would be randomly expected (*p*-value <0.001; see Materials and Methods). This result suggests a tendency for interface residues to co-evolve. Interacting pairs from which both proteins have an nsSNP overlapping a predicted binding site are given in Table S4.

Table 1 – Functionally annotated protein sites that coincide with predicted interaction sites.

Protein/Gene name	TAIR/UNIPROT	Amino acids/Mutation	Mutagenesis Effect or Region Annotation	Reference	Predicted site
Acyl-CoA binding protein 5 (ACBP5)	AT5G27630/Q8RWD9	46, 53, 75 and 94 / L->Q, Q->A, K->A, F->A	Reduction of oleoyl-CoA-binding	[49]	41 to 48; 51 to 58; 71 to 83; 89 to 94
AFPH2(NINJA)	AT4G28910/Q9SV55	7 to 17	Necessary for the interaction with TOPLESS	[50]	16 to 23
		322 to 425	Necessary for the interaction with the JAZ proteins	[50]	344 to 351; 353 to 360
AtBRE1(HUB1)	AT2G44950/Q8RXD6	712 to 878 / Missing in mutant hub1-1/ang4-1	Loss of function	[51]	859 to 869
AtCAND1(CAND1)	AT2G02560/Q8L5Y6	1069 / G->D	Reduced response to auxin	[52]	1062 to 1069
CXIP1(GRXS14)	AT3G54900/Q84Y95	133 to 137 / SNWPT->AAAAA	Loss of CAX1 activation	[22]	125 to 136
CONSTANS(CO)	AT5G15840/Q39057	96 to 98 / Missing in mutant co-1	Late-flowering under long day condition	[53]	93 to 100
IAA3(SHY2)	AT1G04240/Q38822	67 and 69 / G->E and P->S	Affects auxin-related developmental processes	[23]	59 to 69
IAA7(AXR2)	AT3G23050/Q38825	87 / P->S	Affects auxin-related developmental processes	[54]	77 to 95
IAA19(MSG2)	AT3G15540/O24409	3, 75 and 76 / G -> R, P -> L and P -> L	Affects auxin-related developmental processes	[55]	67 to 74
PHABULOSA(ATHB-14)	AT2G34710/O04291	202 / G->D	Transformation of abaxial leaf fates into adaxial leaf fates	[56]	198 to 204
TGA1(BZIP47)	AT5G65210/Q39237	260 / C->N	Gain of interaction with NPR1	[57]	257 to 264
TIFY 10A(JAZ1)	AT1G19180/Q9LMA8	202 to 228 / region missing in mutant jaz1delta3A	Dominant mutation that confers jasmonate insensitivity	[58]	213 to 220
TIFY 6B(JAZ3)	AT3G17860/Q9LVI4	299 to 312 / VALPLARKASLARF->GKKQSQRPDTTFAI	Dominant mutation that confers jasmonate insensitivity	[59]	309 to 318
TOPLESS(TLP)	AT1G15750/Q94AI7	176 / K-> M	Temperature sensitive gain of function	[60]	171 to 178
YABBY 4(YAB4)	AT1G23420/Q9LDT3	147 / K->KLYWSR	Reduced development of the ovule outer integument	[61]	126 to 166
ZEITLUPE(ZTL)	AT5G57360/Q94BT6	200 and 213 / L->A, L->A	No ZTL-ASK1 complex formation	[20]	208 to 220

Putative molecular mechanisms underlying effects of amino acid mutagenesis

A major application of our predictions is to provide sites that can be targeted by mutagenesis to change the interaction specificity of a protein, and to provide putative explanations for observed phenotypic changes upon mutations in terms of changes in interaction specificity. To assess the

usefulness of our data towards these goals, we compared our predictions with available results from experimental mutagenesis experiments and their effects on molecular functions and biological processes (seeMaterials and Methods). The experimentally annotated mutagenesis sites considered here, in general involve residues that are located in functional sites, which in certain number of cases corresponds to protein-protein interaction sites. Hence, one would expect a tendency for the predicted binding sites to coincide with such annotated sites. This was indeed the case: out of 38 proteins for which mutagenesis data is available and for which we predicted the interaction site, for 16 there is at least one mutation site that coincides with a predicted binding site (Table 1).

By analysing details of available annotation for those cases where a predicted binding site coincides with an experimentally annotated mutagenesis site, we found that some of them are indeed involved in protein interactions, whereas for others this is not known but our results provide evidence for such role. For example, in the protein ZEITLUPE (ZTL, AT5G57360), alanine mutagenesis of the residues 200 or 213 located in the F-box domain eliminates the interaction with ASK1 (AT1G75950), in the yeast-two-hybrid system and *in vitro* [20]. Accordingly, for ZEITLUPE, the stretch of residues from 208 to 220 is predicted as interaction site for binding with ASK2 (AT5G42190) and ASK4 (AT1G20140). This leads to the hypothesis that mutation on the F-box domain of ZEITLUPE, specifically in residue Leu213, would not only disrupt its interaction with ASK1, but also with other SKP1-like proteins [21], such as ASK2 and ASK4 (Figure 4, A–B).

A similar case is obtained by analysing available annotation of the protein CXIP1 (GRXS14, AT3G54900), which is thought to activate CAX1 (AT2G38170) through a direct interaction. In CXIP1, alanine mutagenesis of two highly conserved motifs (SNWPT, residues from 133 to 137; and CGFS, residue from 97 to 100) has been shown to lead to loss of ability to activate CAX1, presumably by abolishing the interaction between these two proteins [22]. For CXIP1, we predicted as binding site the stretch of residues from 125 to 136, which overlaps one of the mutation positions. Although CAX1 is not represented in the Arabidopsis interactome data, four other interaction partners for CXIP1 have been identified; i.e. AT5G09830, AT3G50780, AT1G70410 and TCP13 (AT3G02150). We predict that the interaction of CXIP1 with these proteins may also be mediated by the same SNWPT motif (Figure 4, C–D).

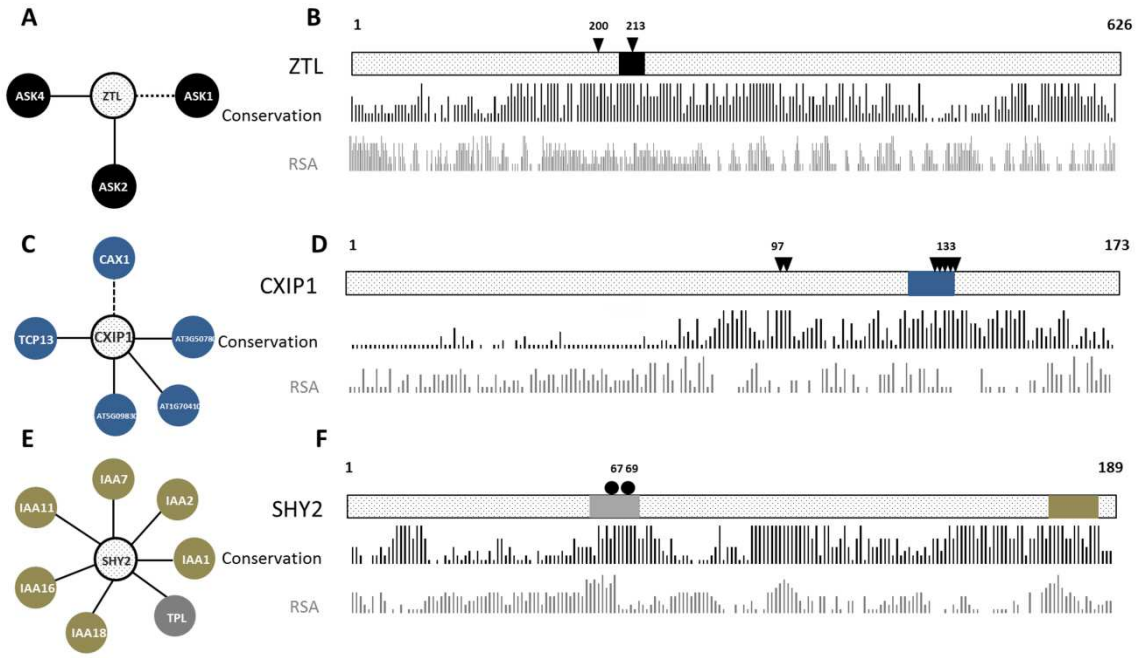


Figure 4: Putative molecular mechanisms underlying effects of amino acid mutagenesis. A, C and E show the interacting partners of the proteins ZTL, CXIP1 and SHY2, respectively (interactions shown as dashed lines are not covered in the Arabidopsis Interactome data). B, D and F show a schematic representation of the sequences of the three proteins, including predicted binding sites (coloured box, using same colour as the proteins predicted to bind to it), mutagenesis sites (triangles for experimental mutagenesis sites, circles for naturally occurring sequence variants) and their positions, and residue surface accessibility (RSA) and conservation (bar plots) as predicted based on the sequence. A–B, in the protein ZTL, alanine mutagenesis of the residues 200 and 213 independently eliminate the interaction with ASK1; for ZTL, the stretch of residues from 208 to 220 is predicted as interaction site for binding with ASK2 and ASK4. This leads to the hypothesis that mutation on ZTP, specifically on the residue Leu213, would not only disrupt its interaction with ASK1, but also with other SKP1-like proteins, such as ASK2 and ASK4. C–D, In CXIP1, alanine mutagenesis of two highly conserved motifs (residues from 133 to 137; and residues from 97 to 100) leads to loss of ability to activate CAX1. For CXIP1, the stretch of residues from 125 to 136 was predicted as binding site, which overlaps the mutated motif SNWPT. The interaction of CXIP1 and the other interacting partners identified in the Arabidopsis interactome, i.e. AT5G09830, AT3G50780, AT1G70410 and TCP13 (AT3G02150), may also be mediated by the same motif. E–F, in the sequence of SHY2, three motifs were predicted as binding sites. The first (residues from 59 to 69; represented in grey) overlaps the position of two naturally occurring mutations (residues 67 and 69) and is predicted to be responsible for binding of TOPLESS (TPL, AT5G27030). A second motif (residues from 180 to 187; represented in brown) is predicted to be responsible for the interactions of SHY2 with six other IAA proteins. This leads to the hypothesis that two known mutations disrupt the interaction of SHY2 with TPL, but the same mutations do not impede its interaction with other IAA proteins.

Additionally, analysis of available mutagenesis data indicates a number of cases in which mutations are known to affect certain phenotypes, but the molecular mechanism behind this is unknown. Our predictions, together with the Arabidopsis interactome, allow us to generate hypotheses for these unknown mechanisms, which could in principle be experimentally tested. For example, for two naturally occurring mutations in SHY2 (IAA3, AT1G04240) the phenotypic effects have been identified: *shy2-2*, in which a proline in position 69 is mutated to a serine; and *shy2-3*, in which a glycine in position 67 is mutated to a glutamic acid. Although both mutations are known to interfere with auxin-related developmental processes, i.e. root growth, gravitropism and lateral root formation [23], the molecular mechanisms underlying these changes are unknown. In the SHY2 sequence, we predicted as binding site three motifs. One of these, the stretch of residues from 59 to 69, overlaps the position of the two known mutations and is predicted to be responsible for binding of TOPLESS (TPL, AT5G27030). A second motif (residues from 180 to 187) is predicted to be responsible for interaction of SHY2 with six other IAA [24] proteins: IAA1 (AT4G14560), IAA2 (AT3G23030), IAA7 (AT3G23050), IAA11 (AT4G28640), IAA16 (AT3G04730) and IAA18 (AT1G51950). This leads to the hypothesis that mutations in positions 67 and 69 of SHY2 may affect its ability to interact with TOPLESS, but the same mutations do not impede the interaction with other IAA proteins (Figure 4, E–F). Note that the predicted binding site in SHY2 occurs in a region (IAA domain II) which is known to be important for the interaction between IAA proteins and F-box containing proteins [25].

Gene duplication and protein-protein interaction network evolution

Gene duplication is a major driving force of evolutionary novelty [10]. Because of redundancy immediately after the duplication event, the selective pressure on one of the two copies might be relaxed, both on its *cis*-regulatory elements and its coding sequence. In the latter case, mutations in protein-protein binding sites may either abolish existing interactions or create new interaction sites. These mutations lead to interaction rewiring as one of the mechanisms for functionalization [26]. Here, to assess to which extent mutations in protein-protein binding sites reflect functional divergence, we used our predictions to examine the sequences of 32 paralogous Arabidopsis protein pairs that have previously been classified as having either “no”, “low”, or “high” functional divergence[27] based on examination of knock-out phenotypes.

For the examined paralogous pairs, the sequence identity of the predicted binding sites was better able than the identity of the whole protein sequence to distinguish the three functional divergence groups (Figure 5; Materials and Methods; Table S5). The weak discriminatory power observed by comparing the three density functions for “whole protein sequence identities” (Figure 5A) means that comparing full-length sequence identity gives only a weak

indication whether two paralogs are likely to be functionally redundant or functionally divergent. In contrast, the differences among the density functions for the "binding site sequence identities" (Figure 5B) suggests that we may predict the degree of functional divergence based on small sequence changes in the binding sites of paralogous pairs.

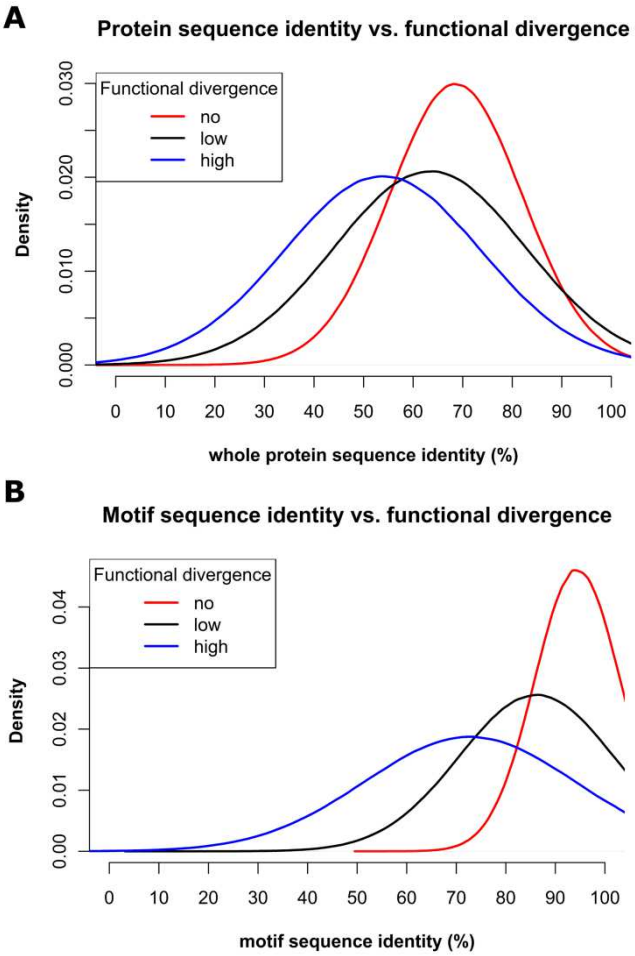


Figure 5: Binding sites contain signal about functional divergence. Distributions of sequence identity values are shown for paralogous pairs classified as having "no" (red), "low" (black) or "high" (blue) functional divergence. The x-axis represents the sequence identity of paralogous pairs. For each paralogous pair, the sequence identity was calculated using either (A) the whole protein sequences, or (B) just the sequence of predicted binding sites. The better separation between the curves for no functional divergence vs. high functional divergence when using predicted interaction sites indicates that these contain signal related to functional divergence.

The potential for exploiting the sequence of binding sites towards predictions of functional divergence may be illustrated by examining the two paralogs FT (AT1G65480) and TFL1 (AT5G03840). Both genes mediate signals for floral transition in an antagonistic manner: whilst the knockout mutant of FT strongly induces late flowering, the knockout mutant of TFL1 induces early flowering [28]. Based on the overall sequence identity (55%) the pair FT/TFL1 would be classified as non-diverged; however, when using the binding site sequence identity

(70%) its most likely classification is “high functional divergence”: the curve for ‘no functional divergence’ has the highest density at 55% for overall sequence identity, but the lowest density at 70% for motif sequence identity (Figure 5). Thus, despite the high overall sequence identity of FT/TFL1, we could correctly infer that the pair shows high functional divergence.

Concluding remarks

Efficient bioinformatics strategies are crucial to retrieve information encoded in biological networks, in particular to support the formulation of hypotheses on evolutionary processes and molecular mechanisms linking genotype to phenotype. Here, we addressed the challenge of locating, at a large scale, protein binding sites in the Arabidopsis proteins. For this task, we defined a strategy that exploits information encoded in the Arabidopsis interactome and the sequences coding for the interacting proteins. Our sequence- and interactome-based approach enabled the prediction of binding motifs in 985 (36%) of the proteins represented in the interactome. Although this number represents only a small percentage of all Arabidopsis proteins, it is much higher than would be expected from methods that rely on protein structure information. One possible way to achieve higher coverage would be by using a different set of parameters controlling the thresholds of evolutionary conservation and surfaces accessibility of the motifs. Alternatively, predictions based on additional protein-protein interaction datasets [29], [30], [31], [32] could complement the current set of predictions, as will future extensions of the Arabidopsis interactome data. In addition, we recently also developed an extension of the SLIDER algorithm which obtains a much higher coverage of a given network of proteins (Boyen *et al.*, submitted to *Trans Comp Biol Bioinf*) although this does not yet use the biological information sources applied in the current study.

We used our predictions to investigate evolutionary aspects of binding site variability. By assessing the frequency of synonymous and non-synonymous SNPs either in the whole protein sequence or only in the predicted motifs, we found that, overall, our predicted sites are under stronger evolutionary constraints than the rest of the protein. Additionally, we identified non-synonymous SNPs that may be correlated with changes in the protein interaction specificity between different Arabidopsis ecotypes.

Previously, we employed sequence-based approaches [33] to mine binding motifs from the interaction network of transcription factors [29] belonging to the MADS domain protein family [34]. Although the approach used in that work is not applicable to a large interactome due to computational complexity of the algorithm, these results were used to experimentally change the interaction specificity of several MADS domain proteins. This provided insight into mechanisms underlying sub- or neo-functionalization among members of the MADS box family. Here, to corroborate our proteome-scale predictions we used available mutagenesis data

(Table 1) to form testable hypotheses for the molecular mechanisms underlying effects of known mutations on several proteins (Figure 4). Our predicted interaction sites are available at www.ab.wur.nl/sliderbio and can be used to pinpoint residues which should be mutated in order to interfere with specific interactions, or to interpret the results of obtained phenotypic changes upon mutations in a molecular and mechanistic way. They also enable to perform large scale studies on the effects of various types of naturally occurring sequence variation on protein interactions, similar to what we recently demonstrated for the MADS domain protein family [35].

It has been debated whether constraints placed on binding sites play a major role in functional divergence [36], when compared to constraints placed on *cis*-elements. Here, Arabidopsis paralogous pairs that have previously been classified, based on morphological changes observed upon mutation, into functional divergence groups [27]were analysed. From our analysis, it seems that the sequence identity calculated over the whole sequence does not contain a lot of signal that explain the observed divergence (Figure 5A). This is in agreement with the findings of [11], in which the correlation between selective pressure on the whole sequence and the functional divergence was assessed. However, when we analysed only the sequence region covered by binding sites (Figure 5B), we found a stronger correlation between functional divergence and selective pressure. Obviously, this does not mean that non-coding sequence divergence (in particular via its effect on gene expression) would not be important for functional divergence, but it demonstrates the importance of coding sequence variation as an additional factor. These examples set the stage for future investigation of the correlation between sequence divergence and phenotypic divergence.

Materials and Methods

Protein-protein interactions and sequence data

The *Homo sapiens* (human) and *Saccharomyces cerevisiae* (yeast) protein-protein interaction data used in this work are described in [37]. The *Arabidopsis thaliana*interaction data were obtained from the recently published Arabidopsis interactome map[7]. The sequences of human, yeast and Arabidopsis proteins were retrieved, respectively, from the UniProt [38], Saccharomyces Genome [39] and TAIR [40] databases (see Table S1).

Mapping protein interacting pairs to known complex structures

One of our assessment procedures aims to verify whether the predicted motifs are located in the protein-protein interface, which is a straightforward task when the structure of the complex is available. However, few complex structures deposited in the PDB correspond to the proteins listed in PPI data used in this work. To overcome such a lack of structural information, we used

a strategy to assign sequences to known protein structures based on homology. To link a query sequence to a target sequence with a known 3D structure, we used PSI-BLAST [41] to search against the PDB database under the following conditions: (1) the bit score is higher than 70; (2) the aligned region from the query sequence has a length that corresponds to at least 30% of the query total length; (3) the aligned region from the target sequence has a length that corresponds to at least 30% of the target total length; and (4) the identity of the aligned regions is higher than 40%. Subsequently, we used the sequences and their assigned structures to filter the interacting lists to retain only the interactions for which both proteins link to interacting units of a complex with known structure (e.g. proteins A and B interact, and protein sequence A is assigned to protein structure X chain K, protein sequence B is assigned to protein structure X chain L). The resulting subsets of protein-protein interactions contain for the human, yeast and Arabidopsis, respectively, 539, 263 and 53 interactions among 575, 213 and 67 proteins. We refer to these subsets of the protein-protein interaction networks as structurally mapped datasets (see Table S2).

Identification of interface residues in protein complex structures

After mapping protein sequences to known structures, the interface residues were identified in the complex structures that were assigned to pairs of interacting proteins. To determine these interface residues, we used NACCESS [42] to calculate the residue solvent accessible surface area for all the complexes and for all the unbound proteins. A residue was classified as interface when the solvent accessible surface area calculated in the complex was smaller than the value calculated in the unbound protein. Following the interface residue identification, the protein sequence was aligned with the sequence of its assigned PDB using Clustal [43] and the alignment was used to map residues from the structure to residues in the sequence. In this way, lists of interface residues and non-interface residues of the interacting proteins comprising the structurally mapped datasets were obtained. This data was used to analyse the performance of the various SLIDERBio parameter settings. Note that as input for SLIDERBio itself, only sequence-based information (conservation and predicted surface accessibility) is used.

Implementation of conservation, accessibility and similarity matrix in SLIDERBio

We extended the original SLIDER algorithm by adding filtering steps based on evolutionary conservation and surface accessibility as predicted from protein sequences, and by implementing an approach to define the presence of a motif in a sequence based on a substitution matrix. Below, we describe these adjustments to the algorithm.

Calculating residue conservation scores.

Calculating residue conservation requires three sequential tasks: to select a group of homologous proteins, to align the protein sequence with these homologs, and to quantify the conservation of each residue in the alignment. To select groups of homologs we used OrthoMCL (Version 2.0; [44]) to assign each protein to an OrthoMCL-DB (release 5) group. Next, we used Clustal [43] to align the protein sequence with the sequences of all members of the associated OrthoMCL-DB group. Finally, we used the AL2CO software [45]to obtain a conservation score for each position in the multiple sequence alignments. The AL2CO algorithm performs its calculation in two steps: first amino acid frequencies at each position in the alignment are estimated, and then a score is calculated from these frequencies. We used the methods unweight-frequencies and entropy-based in the first and second step, respectively. To assign a conservation score to each residue in the protein sequence, we used the integer conservation indices resulting from the AL2CO calculation. The AL2CO integer conservation score ranges from 0 to 9, representing low to high conservation, respectively; it is obtained from the entropy-score by a linear scaling (subtracting the minimum value and dividing by the difference between maximum and minimum value) To assess the conservation of a given motif, we use the average of the residue conservation scores over the motif length; only if this average is higher or equal than the conservation threshold, SLIDERBio may consider this motif as a binding site.

Calculating residue solvent accessibility scores.

The relative solvent accessibility (RSA) of an amino acid residue in a protein indicates its level of solvent exposure. To predict the RSA based on protein sequences, we used the SABLE [46] software that predicts whole residue relative RSA scores from sequences alone using a neural network algorithm trained on PDB structures. SABLE outputs an integer value for each residue, ranging from 0 to 9, representing ‘fully buried’ to ‘fully exposed’, respectively. This output is defined as the ratio of solvent-exposed surface area of a residue to the maximum obtainable value of the solvent-exposed surface area for this amino acid, linearly rescaled in a similar way as described above for the conservation score.

Strategy to define motif presence based on substitution matrix.

To quantify the overrepresentation of a given motif in the network, our method verifies in how many sequences that motif is present. Instead of searching for perfect matches, SLIDERBio uses a modified version of the BLOSUM62 similarity table to calculate the “degree of similarity” of a given motif for a protein sequence. In this modified similarity table, a perfect amino acid match has value 1, and a non-perfect match has value ranging from 0 to 1 directly

proportional to the BLOSUM62 score (this linear scaling is performed for each of the rows of the matrix separately). Our method calculates the residue similarity score and it averages the value over the motif length. Only if this average is greater than or equal to the "degree of similarity" threshold, SLIDERBio considers the motif present in the protein sequence.

Quality measures for evaluating predictions of protein-protein binding motifs

To assess the quality of the SLIDERBio results, we defined two measures that use the structures of the proteins in the above-mentioned structurally mapped datasets. Here, the 'Accuracy of predicted motifs' is defined as the number of motifs correctly predicted to be in the interface as a fraction of all motifs predicted to be in protein–protein interface. A motif is said to be in the interface, if at least one of its residues is identified to be in the interface of its assigned complex structure. The 'Coverage of protein-protein interfaces' stands for the number of protein pairs that contain at least one motif mapped to their interface, as fraction of the total number of interacting pairs in the interaction data. Thus, the 'Accuracy of predicted motifs' reflects the predictive power of the algorithm toward finding motifs that are indeed located in the interface, and the 'Coverage of protein-protein interfaces' reflects its predictive power towards finding motifs explaining the largest number of interactions. The overall performance of the predictions was measured via the F-score, which equals 2*'Accuracy of predicted motifs'*'Coverage of protein-protein interfaces'/('Accuracy of predicted motifs '+'Coverage of protein-protein interfaces').

Setting SLIDERBio parameters

For the threshold of the allowed degree of similarity between motif sequence and protein sequence, we tested five different values ([none;0.4;0.5;0.6;0.7], where 'none' stands for not having used the modification). For the thresholds of conservation and residue surface accessibility, we tested six different values ([none;3;4;5;6;7]) each. In total, 180 combinations ($5\times6\times6$) of these values were tested. SLIDERBio predicts a set of N motif pairs. For each combination of parameters, we executed SLIDERBio on the structurally mapped datasets for the three species using the following configuration: length of predicted motif l = 8; number of allowed wildcard-character d = 5; maximum execution time$t$ = 60 minutes; number of predicted motif pairs N = 1,000. We then mapped the resultant motif pairs in the sequence of pairs of interacting proteins, in such a way that each of the interacting proteins contains one of the motifs in the pair. Subsequently, the 'Accuracy of predicted motifs', 'Coverage of protein-protein interfaces' and F-score were calculated for all the results. For the analysis of the complete Arabidopsis Interactome, maximum execution time was set to t = 24 hours.

Mapping predicted motif pairs to protein sequences

We used our method to predict motif pairs that are overrepresented in pairs of interacting proteins, conserved across species, and predicted to be exposed in the protein surface. Each motif can usually be "mapped" to more than one protein sequence. This mapping is performed by searching each of the motifs against all the interacting protein sequences; and considering only those matches that fit both requirements for conservation and surface accessibility (i.e. conservation greater than the conservation threshold and surface accessibility greater than the RSA threshold).

Randomly generated sets of motif pairs

In order to assess the significance of the SLIDERBio results, we created sets of random motif pairs by applying the following strategy: First, we randomly selected a sequence in the input sequence set; next, we randomly sampled from the selected sequence a substring of length *l*, and randomly arranged d wildcard-characters in the substring. The same procedure was repeated to create the second motif in the pair, which resulted in a motif pair. Then, we repeated this step till *N* motif pairs were created. In this way, we created 1,000 sets of *N* motif pairs for each of the structurally mapped datasets (human, yeast and Arabidopsis), using the same set up of parameters controlling the length of the motifs ($l = 8$ and $d = 5$), and the same number of motif pairs ($N = 1{,}000$).

Analysis of single nucleotide polymorphism

SNPs were obtained from the currently available 80 accessions from the Arabidopsis 1001 Genome Project [47]. After mapping to protein coding sequences, non-synonymous SNPs were extracted and their positions were compared with positions of predicted interface residues. To compare the significance of the small overlap between non-synonymous SNPs and binding sites, sets of randomly chosen "SNPs" were generated (with the same number of SNPs per protein as in the experimental data) and their overlap with the binding sites was counted (using 1,000 random trials). To compare the significance of the amount of interactions between proteins with SNPs overlapping predicted interaction sites, we randomly selected the same number of proteins from the interactome and counted their number of interactions (using 1,000 random trials).

Analysis of mutagenesis regions

We retrieved and analysed the field "Experimental info" from the section "Sequence annotation" as deposited in the UniProt database [38]. This describes the effects of mutations of

amino acids on the biological properties of proteins. Out of all the 985 protein with interface residues predicted by SLIDERBio, experimental information was available for 38 proteins.

Gene duplication and Functional divergence analysis

To classify the paralogous pairs as having “no”, “low” or “high” functional divergence, we used data from [27], where the divergence was measured on the basis of morphological consequences observed in null mutants of single genes or pairs of genes. From the obtained list of 492 paralogous pairs, we kept only those pairs from which for at least one of the paralogs interface residues were predicted by SLIDERBio (n = 32). Next, we used Needle [48] to compute the global pairwise alignment and to calculate the “whole protein sequence identity” for each pair (see Table S5). Then, we mapped our predicted motifs to the resultant alignments and calculated the “binding site sequence identity” by comparing only the sequence regions to which motifs were mapped. To avoid bias of motifs mapped in regions with long gaps, we removed from the analysis any motifs that were mapped to gapped regions.

For each functional divergence group (“no”,“low” and “high”), we created two density functions by fitting a normal distribution to the calculated values of either “whole protein sequence identity” or “binding site sequence identity”. Prior to the analyses, we tested the normality of each group of values using Lilliefors test for normality with no significant results (p-values: (0.5, 0.2, 0.2) and (0.1, 0.4, 0.6); for (“no”,“low” and “high” functional divergence) of “whole protein sequence identity” and “motif sequence identity”, respectively), suggesting that the data is normally distributed.

Supporting Information

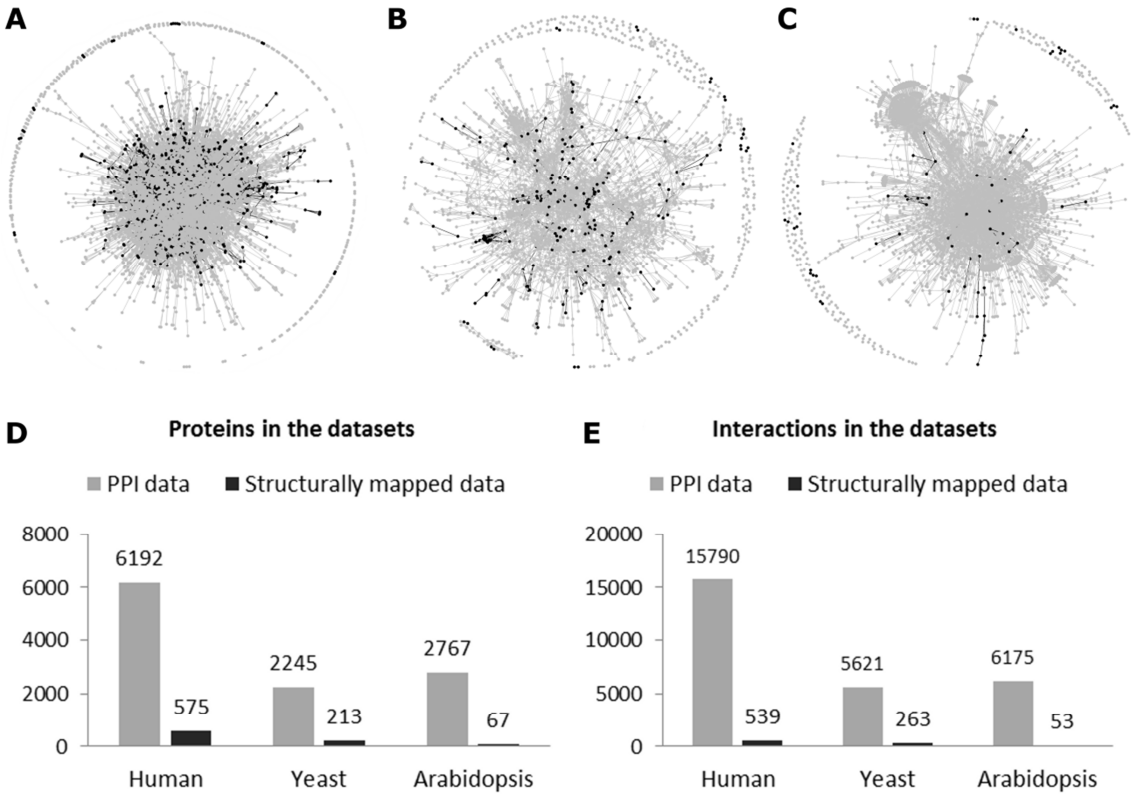


Figure S1:Topological properties of the protein-protein interaction (PPI) networks and their respective structurally mapped subsets. To create the basis for comparison and assessment of our predictions, we used the structures of protein complexes in order to identify residues that are located in the protein interface. Because the number of complex structures mapped to Arabidopsis proteins is low, we used two other datasets from which more structures are available; the human and yeast protein-protein interaction networks. (A–C) Graphical representation of the human (A), yeast (B) and Arabidopsis (C) interactome. Nodes represent proteins, edges represent interactions. (D) number of proteins and (E) number of interactions in the PPI datasets. Black, proteins and interactions from which structures could be mapped; grey, complete PPI data.

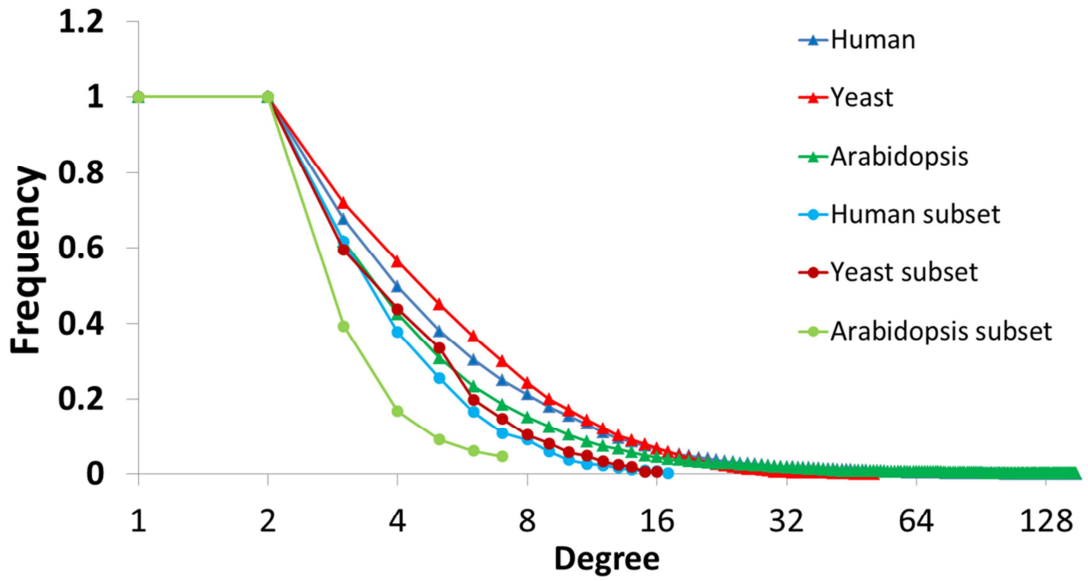


Figure S2: Comparison of the topology of the protein-protein interaction networks and their respective structurally mapped subsets. *x*-axis represents the number of protein partners (degree) and *y*-axis represents the frequency. The Figure allows quantitative comparison of the network composed by the subset of interacting proteins from which structural information is available against the complete set of interactions. By using the degree distributions, we observe that the similarity between the structure mapped subsets for the human and yeast interactomes is high, while the Arabidopsis subset has a quite different degree distribution. In addition, the similarity between the yeast and human structurally mapped datasets and the complete Arabidopsis interactome is higher than the similarity between the Arabidopsis subset and the complete Arabidopsis interactome.

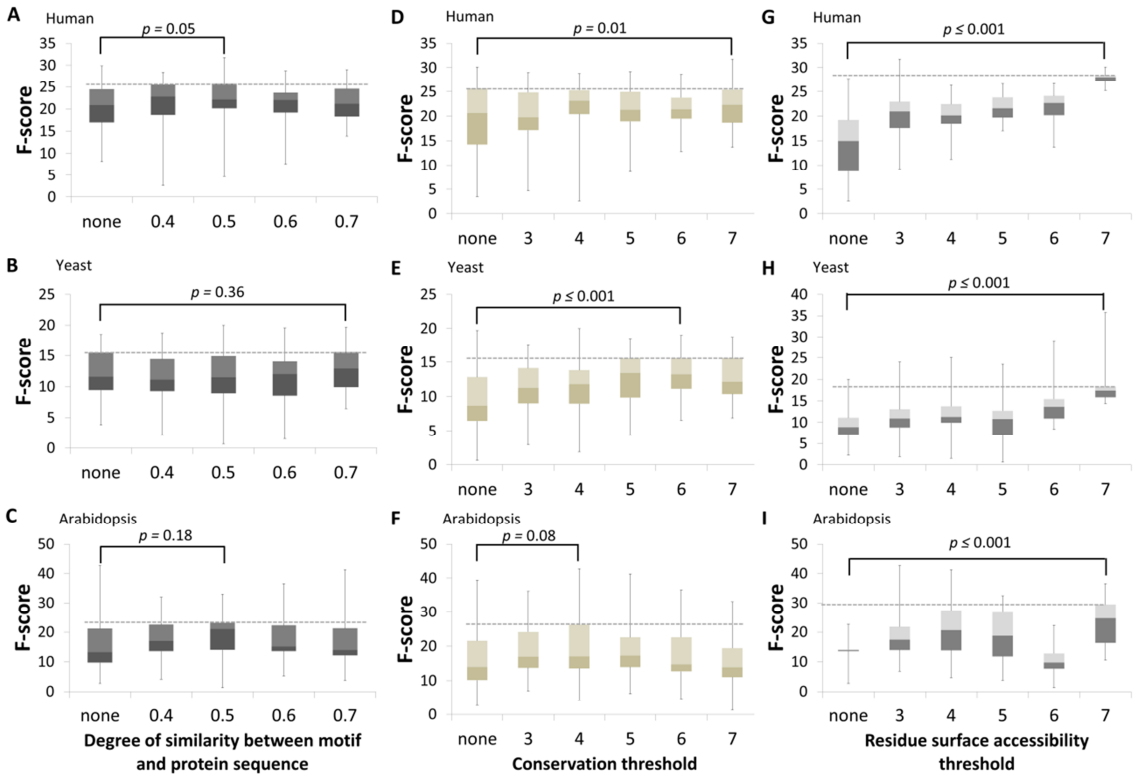
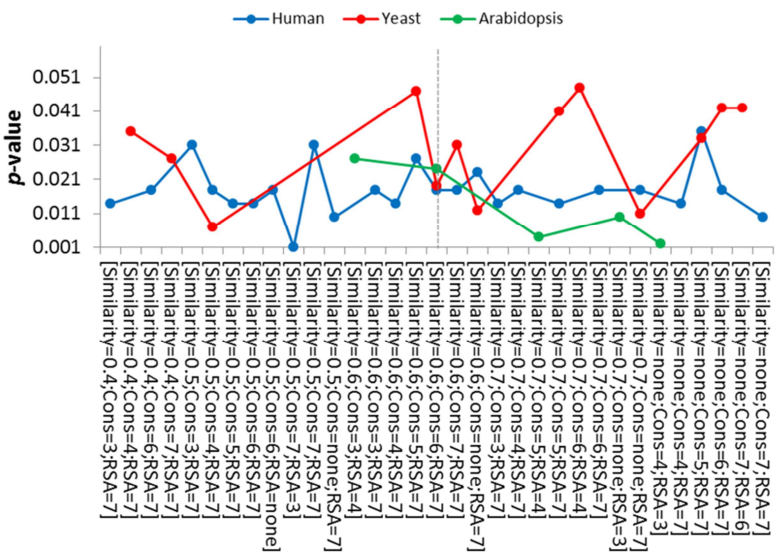


Figure S3: Assessment of the SLIDERBio performance for different values for the thresholds of Degree of similarity, Conservation and surface accessibility. The box plots group the F-score results (*y*-axis) based on each used threshold value for the SLIDERBio parameters: (A,B,C) show the results grouped based on threshold values for the Degree of Similarity between motif and protein sequence; (D,E,F) for the Conservation threshold values; and (G,H,I) for the Residue surface accessibility threshold values. The results for the Human, Yeast and Arabidopsis structurally mapped datasets are shown, respectively, in (A,D,G), (B,E,H) and (C,F,I). The boxes labelled as 'none' contain the F-score results when SLIDERBio did not use the modification in its calculation. The grey horizontal dashed lines touch the boxes in the group that has given greatest 75th percentile. We then tested whether there is statistical difference in the F-score results when SLIDERBio uses or does not use the modification. The figures show the p-value (P) when the results from the group 'none' are compared against the results from the group with greatest F-score 75t^{h}percentile. All *p*-values (*P*) shown in the figures are calculated using a two-tailed paired t-test. At significance level 0.01, we reject the null hypothesis that the means are equal.



<u>Figure S4:</u> Determination of a default set of SLIDERBio parameter values. The figure shows the *p*-values calculated by comparing F-scores obtained from the SLIDERBio results against those from random results. *y*-axis represents the p-value; *x*-axis indicates which combination of parameters has been used. For legibility, only results for which the p-value is less than 0.05 are shown. The vertical dashed grey line indicates the single parameter setting that showed *p*-values less than 0.05 simultaneously for all the three structurally mapped dataset. This combination of parameters [Degree of similarity = 0.6; Conservation = 6; Surface accessibility = 7] is used to predict binding motifs on the full Arabidopsis interactome.

Table S1. Human, yeast and Arabidopsis protein-protein interaction networks used in this work. (XLSX available online)

Table S2. Structures of protein complexes mapped to sequences of interacting proteins. (XLSX available online)

Table S3. Predicted interaction motifs for Arabidopsis proteins. (XLSX)

Table S4. List of interacting proteins in which a nsSNP overlaps the binding site of both proteins. (XLSX available online)

Table S5. Functional divergence classification and sequence similarity analysis of paralogous pairs with predicted motifs. (XLSX available online)

Author Contributions

Conceived and designed the experiments: FLV PB FN ADJvD. Performed the experiments: FLV PB ADJvD. Analyzed the data: FLV PB ADJvD. Wrote the paper: FLV PB FN ADJvD.

Funding: This work was supported by an Netherlands Organisation for Scientific Research (NWO) VENI grant (863.08.027) to ADJvD, the SYSFLO Marie Curie Initial Training Network (FLV), and a PhD grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen) to PB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Janin J, Rodier F (1995) Protein-protein interaction at crystal contacts. Proteins 23: 580–587. doi: 10. 1002/prot.340230413
2. Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. J Mol Biol 280: 1–9. doi: 10. 1006/jmbi.1998. 1843
3. Moreira IS, Fernandes PA, Ramos MJ (2007) Hot spots–a review of the protein-protein interface determinant amino-acid residues. Proteins 68: 803–812. doi: 10. 1002/prot.21396
4. de Vries SJ, Bonvin AM (2008) How proteins get in touch: interface prediction in the study of biomolecular complexes. Curr Protein Pept Sci 9: 394–406. doi: 10. 2174/138920308785132712
5. Morsy M, Gouthu S, Orchard S, Thorneycroft D, Harper JF, et al. (2008) Charting plant interactomes: possibilities and challenges. Trends Plant Sci 13: 183–191. doi: 10. 1016/j.tplants.2008. 01. 006
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235–242. doi: 10. 1093/nar/28. 1. 235
7. Dreze M, Carvunis AR, Charloteaux B, Galli M, Pevzner SJ, et al. (2011) Evidence for network evolution in an Arabidopsis interactome map. Science 333: 601–607. doi: 10. 1126/science.1203877
8. Tang H, Bowers JE, Wang X, Ming R, Alam M, et al. (2008) Synteny and collinearity in plant genomes. Science 320: 486–488. doi: 10. 1126/science.1153917
9. Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. Nat Rev Genet 10: 725–732. doi: 10. 1038/nrg2600
10. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155. doi: 10. 1126/science.290. 5494. 1151
11. Hakes L, Lovell SC, Oliver SG, Robertson DL (2007) Specificity in protein interactions and its relationship with sequence diversity and coevolution. Proc Natl Acad Sci U S A 104: 7999–8004. doi: 10. 1073/pnas.0609962104
12. Neduva V, Russell RB (2005) Linear motifs: evolutionary interaction switches. FEBS Lett 579: 3342–3345. doi: 10. 1016/j.febslet.2005. 04. 005
13. Valdar WS, Thornton JM (2001) Protein-protein interfaces: analysis of amino acid conservation in homodimers. Proteins 42: 108–124. doi: 10. 1002/1097-0134(20010101)42:1<108::aid-prot110>3. 0. co;2-o
14. Zhang QC, Petrey D, Norel R, Honig BH (2010) Protein interface conservation across structure space. Proceedings of the National Academy of Sciences of the United States of America 107: 10896–10901. doi: 10. 1073/pnas.1005894107
15. Caffrey DR, Somaroo S, Hughes JD, Mintseris J, Huang ES (2004) Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci 13: 190–202. doi: 10. 1110/ps.03323604
16. Boyen P, Van Dyck D, Neven F, van Ham RC, van Dijk AD (2011) SLIDER: A Generic Metaheuristic for the Discovery of Correlated Motifs in Protein-Protein Interaction Networks. IEEE/ACM Trans Comput Biol Bioinform 8 (5) 1344–1357. doi: 10. 1109/tcbb.2011. 17
17. Eddy SR (2004) Where did the BLOSUM62 alignment score matrix come from? Nature biotechnology 22: 1035–1036. doi: 10. 1038/nbt0804-1035
18. Mika S, Rost B (2006) Protein-protein interactions more conserved within species than across species. PLoS Comput Biol 2: e79. doi: 10. 1371/journal.pcbi.0020079
19. Pazos F, Valencia A (2008) Protein co-evolution, co-adaptation and interactions. The EMBO journal 27: 2648–2655. doi: 10. 1038/emboj.2008. 189

20. Han L, Mason M, Risseeuw EP, Crosby WL, Somers DE (2004) Formation of an SCF(ZTL) complex is required for proper regulation of circadian timing. Plant J 40: 291–301. doi: 10. 1111/j.1365-313x.2004. 02207. x
21. Zhao D, Ni W, Feng B, Han T, Petrasek MG, et al. (2003) Members of the Arabidopsis-SKP1-like gene family exhibit a variety of expression patterns and may play diverse roles in Arabidopsis. Plant Physiol 133: 203–217. doi: 10. 1104/pp.103. 024703
22. Cheng NH, Hirschi KD (2003) Cloning and characterization of CXIP1, a novel PICOT domain-containing Arabidopsis protein that associates with CAX1. J Biol Chem 278: 6503–6509. doi: 10. 1074/jbc.m210883200
23. Tian Q, Reed JW (1999) Control of auxin-regulated root development by the Arabidopsis thaliana SHY2/IAA3 gene. Development 126: 711–721. doi: 10. 1105/tpc.010283
24. Reed JW (2001) Roles and activities of Aux/IAA proteins in Arabidopsis. Trends Plant Sci 6: 420–425. doi: 10. 1016/s1360-1385(01)02042-8
25. Kepinski S, Leyser O (2005) The Arabidopsis F-box protein TIR1 is an auxin receptor. Nature 435: 446–451. doi: 10. 1038/nature03542
26. Sun MG, Kim PM (2011) Evolution of biological interaction networks: from models to real data. Genome Biol 12: 235. doi: 10. 1186/gb-2011-12-12-235
27. Hanada K, Kuromori T, Myouga F, Toyoda T, Shinozaki K (2009) Increased expression and protein divergence in duplicate genes is associated with morphological diversification. PLoS genetics 5: e1000781. doi: 10. 1371/journal.pgen.1000781
28. Kobayashi Y, Kaya H, Goto K, Iwabuchi M, Araki T (1999) A pair of related genes with antagonistic roles in mediating flowering signals. Science 286: 1960–1962. doi: 10. 1126/science.286. 5446. 1960
29. de Folter S, Immink RG, Kieffer M, Parenicova L, Henz SR, et al. (2005) Comprehensive interaction map of the Arabidopsis MADS Box transcription factors. Plant Cell 17: 1424–1433. doi: 10. 1105/tpc.105. 031831
30. Causier B, Ashworth M, Guo W, Davies B (2012) The TOPLESS interactome: a framework for gene repression in Arabidopsis. Plant Physiol 158: 423–438. doi: 10. 1104/pp.111. 186999
31. Van Leene J, Boruc J, De Jaeger G, Russinova E, De Veylder L (2011) A kaleidoscopic view of the Arabidopsis core cell cycle interactome. Trends Plant Sci 16: 141–150. doi: 10. 1016/j.tplants.2010. 12. 004
32. Wang C, Marshall A, Zhang DB, Wilson ZA (2012) ANAP: an integrated knowledge base for Arabidopsis protein interaction network analysis. Plant Physiol 158 (4)1523–1533. doi: 10. 1104/pp.111. 192203
33. van Dijk AD, ter Braak CJ, Immink RG, Angenent GC, van Ham RC (2008) Predicting and understanding transcription factor interactions based on sequence level determinants of combinatorial control. Bioinformatics 24: 26–33. doi: 10. 1093/bioinformatics/btm539
34. van Dijk AD, Morabito G, Fiers M, van Ham RC, Angenent GC, et al. (2010) Sequence motifs in MADS transcription factors responsible for specificity and diversification of protein-protein interaction. PLoS Comput Biol 6: e1001017. doi: 10. 1371/journal.pcbi.1001017
35. Severing EI, van Dijk AD, Morabito G, Busscher-Lange J, Immink RG, et al. (2012) Predicting the Impact of Alternative Splicing on Plant MADS Domain Protein Function. PLoS One 7: e30524. doi: 10. 1371/journal.pone.0030524
36. Guharoy M, Chakrabarti P (2005) Conservation and relative importance of residues across protein-protein interfaces. Proc Natl Acad Sci U S A 102: 15447–15452. doi: 10. 1073/pnas.0505425102
37. Yu J, Guo M, Needham CJ, Huang Y, Cai L, et al. (2010) Simple sequence-based kernels do not predict protein-protein interactions. Bioinformatics 26: 2610–2614. doi: 10. 1093/bioinformatics/btq483

38. Consortium TU (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic acids research 38: D142–148. doi: 10. 1093/nar/gkp846
39. Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, et al. (1998) SGD: Saccharomyces Genome Database. Nucleic acids research 26: 73–79. doi: 10. 1093/nar/26. 1. 73
40. Poole RL (2007) The TAIR database. Methods in molecular biology 406: 179–212. doi: 10. 1007/978-1-59745-535-0_8
41. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25: 3389–3402. doi: 10. 1093/nar/25. 17. 3389
42. Hubbard SJ TJ (1993) 'NACCESS', Computer Program. London: Department Molecular Biology University College.
43. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2. 0. Bioinformatics 23: 2947–2948. doi: 10. 1093/bioinformatics/btm404
44. Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. Nucleic acids research 34: D363–368. doi: 10. 1093/nar/gkj123
45. Pei J, Grishin NV (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. Bioinformatics 17: 700–712. doi: 10. 1093/bioinformatics/17. 8. 700
46. Adamczak R, Porollo A, Meller J (2005) Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 59: 467–475. doi: 10. 1002/prot.20441
47. Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, et al. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet 43: 956–963. doi: 10. 1038/ng.911
48. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of molecular biology 48: 443–453. doi: 10. 1016/0022-2836(70)90057-4
49. Leung KC, Li HY, Mishra G, Chye ML (2004) ACBP4 and ACBP5, novel Arabidopsis acyl-CoA-binding proteins with kelch motifs that bind oleoyl-CoA. Plant Mol Biol 55: 297–309. doi: 10. 1007/s11103-004-0642-z
50. Pauwels L, Barbero GF, Geerinck J, Tilleman S, Grunewald W, et al. (2010) NINJA connects the co-repressor TOPLESS to jasmonate signalling. Nature 464: 788–791. doi: 10. 1038/nature08854
51. Fleury D, Himanen K, Cnops G, Nelissen H, Boccardi TM, et al. (2007) The Arabidopsis thaliana homolog of yeast BRE1 has a function in cell cycle regulation during early leaf and root growth. Plant Cell 19: 417–432. doi: 10. 1105/tpc.106. 041319
52. Chuang HW, Zhang W, Gray WM (2004) Arabidopsis ETA2, an apparent ortholog of the human cullin-interacting protein CAND1, is required for auxin responses mediated by the SCF(TIR1) ubiquitin ligase. Plant Cell 16: 1883–1897. doi: 10. 1105/tpc.021923
53. Laubinger S, Marchal V, Le Gourrierec J, Wenkel S, Adrian J, et al. (2006) Arabidopsis SPA proteins regulate photoperiodic flowering and interact with the floral inducer CONSTANS to regulate its stability. Development 133: 3213–3222. doi: 10. 1242/dev.02481
54. Liscum E, Reed JW (2002) Genetics of Aux/IAA and ARF action in plant growth and development. Plant Mol Biol 49: 387–400. doi: 10. 1007/978-94-010-0377-3_10
55. Tatematsu K, Kumagai S, Muto H, Sato A, Watahiki MK, et al. (2004) MASSUGU2 encodes Aux/IAA19, an auxin-regulated protein that functions together with the transcriptional activator NPH4/ARF7 to regulate differential growth responses of hypocotyl and formation of lateral roots in Arabidopsis thaliana. Plant Cell 16: 379–393. doi: 10. 1105/tpc.018630

56. McConnell JR, Emery J, Eshed Y, Bao N, Bowman J, et al. (2001) Role of PHABULOSA and PHAVOLUTA in determining radial patterning in shoots. Nature 411: 709–713. doi: 10. 1038/35079635
57. Despres C, Chubak C, Rochon A, Clark R, Bethune T, et al. (2003) The Arabidopsis NPR1 disease resistance protein is a novel cofactor that confers redox regulation of DNA binding activity to the basic domain/leucine zipper transcription factor TGA1. Plant Cell 15: 2181–2191. doi: 10. 1105/tpc.012849
58. Thines B, Katsir L, Melotto M, Niu Y, Mandaokar A, et al. (2007) JAZ repressor proteins are targets of the SCF(COI1) complex during jasmonate signalling. Nature 448: 661–665. doi: 10. 1038/nature05960
59. Chini A, Fonseca S, Fernandez G, Adie B, Chico JM, et al. (2007) The JAZ family of repressors is the missing link in jasmonate signalling. Nature 448: 666–671. doi: 10. 1038/nature06006
60. Long JA, Ohno C, Smith ZR, Meyerowitz EM (2006) TOPLESS regulates apical embryonic fate in Arabidopsis. Science 312: 1520–1523. doi: 10. 1126/science.1123841
61. Villanueva JM, Broadhvest J, Hauser BA, Meister RJ, Schneitz K, et al. (1999) INNER NO OUTER regulates abaxial- adaxial patterning in Arabidopsis ovules. Genes Dev 13: 3160–3169. doi: 10. 1101/gad.13. 23. 3160

Chapter 5

Concluding remarks

Felipe Leal Valentim

Lessons learned on the road from networks to predictive models

One goal of systems biology is to provide systems-wide snapshots in which the relationships and interactions between the components of a system are comprehensively represented. In the case of gene regulatory networks (GRNs), technological progress in 'omics' approaches (such as RNA-seq, ChIP-seq, DNase-seq, MNase-seq and other new emerging high throughput technologies), and advances in computational methods, in combination with increasing biological knowledge, enabled us to elucidate systems-wide snapshots of complex GRNs underling key plant developmental processes (reviewed in [1]). In order to represent the static snapshots as manipulable models, one must make use of existent modelling frameworks [2]. In case of quantitative modelling, it is important to generate also accurate experimental data that can be used as input in the modelling framework in order to fit the model parameters. Accurate data would mean high spatial-temporal resolution and also high quantitative resolution. However, in spite of improvements and advancements, often modelling of GRNs has to deal with 1) incomplete, inaccurate or contradictory representations of GRNs or 2) experimental data with unsatisfactory resolution. Even if the snapshots and the associate data are satisfactory, modelling of GRNs has also to deal with 3) limitations of the mathematical framework.

In Chapter 2, we present a quantitative model that represents the dynamics of the GRN of flowering time control. For that GRN model, experimental data renders a clear, comprehensive and accurate snapshot of the interactions and regulatory relationships between the involved molecules. But because of the data resolution, we 1) ignored important aspects of molecular localization in cells and tissues, 2) assumed that control of expression is at the transcriptional level only, and 3) assumed that the dependence of translation and transcription rates on protein and RNA concentration, respectively, is strictly monotonic. These are common limitations when modelling gene regulatory networks [3]. In addition, there are also limitations on the mathematical framework. Our proposed model for the GRN of flowering time control is based on ordinary differential equations (ODEs). Most of the models based on ODEs neglect spatial aspects, i.e. it is commonly assumed that gene products move freely within the cell or even between cells. However, it is likely that space and diffusion are important aspects in the dynamics of genetic regulatory systems, which are not incorporated in our model. In addition, our model uses Michaelis-Menten functions. These use saturation and threshold kinetics, as well as degradation parameters, to offer a realistic approximation of enzymatic processes. The Michaelis-Menten functions are usually used to model GRNs because they provide the complexity necessary to represent non-trivial behaviours as observed in gene regulation, but the hypotheses underlying these approximations are rarely shown to be satisfying [4].

Nevertheless, a good fit between predicted and observed gene expressions, as well as a good fit between predicted and observed flowering time, supported the credibility of our model and its dynamic properties. A challenge for the near future will be to increase the spatial-temporal and quantitative resolutions of the experiments (e.g. gene expression, protein abundance, subcellular localization) that characterize the genes in the current networks. The current limitations on data resolution and mathematical framework make current quantitative models suitable for studying the expression dynamics of established regulatory relationships, but much more could be achieved; e.g. unveiling molecular mechanisms underlying the observed patterns of expression.

Our proposed GRN model is mainly focused on transcription factor (TFs), TF–TF and TF–DNA interactions, but we modelled also the relationship between TF expression and the phenotype. This was done by 1) assessing the model predictions of flowering time phenotype for different mutant backgrounds, followed by 2) adjusting the range of model parameters accordingly; and finally by 3) performing new rounds of fitting the equations to the data until improvement was no longer observed. Thus, we obtained a model that not only has a good fit of expression data, but also predicts the flowering time for perturbed systems with good accuracy. It is not often reported that phenotypic information obtained from diverse genetic backgrounds are included in the fitting strategy. Moreover, we attempted to represent the transport of the protein FLOWERING LOCUS T (FT) by naively including a parameter that represents the delay between the timing of FT translation (in the leaves) and FT activity (in the meristem). A future improvement in the model will be to represent the transport of FT more realistically. Movement of FT protein occurs through the phloem system probably by a passive transport mechanism; thus the dynamics of this transport can be represented by e.g. modelling FT diffusion from cell to cell.

Our understanding on how TFs regulate gene expression [5] remains far from complete. For our model, we assumed that MADS-box proteins perform their roles as dimers. The interactions represented in our model are based on recent and comprehensive body of experimental evidences. However, breakthrough advancements in characterizing and identifying higher order TF complexes (such as quartets [6]) as well as a better understanding of the many modes of transcriptional and post-transcriptional regulation playing a role in flowering time (such as microRNAs, movable factors, hormones, chromatin modifying proteins, and alternative splicing) offer exciting possibilities for improving our model.

The inclusion of this information will not only depend on increasing the spatial-temporal and quantitative resolutions of the data used in the modelling framework, but also on advances in the mathematical tools.

Future avenues for extending our model also include integrating the effect of environmental signals on the expression of the flowering time genes. For example, recently, the regulatory relationship between the genes *FLOWERING LOCUS T (FT)* and *SHORT VEGETATIVE PHASE (SVP)* was modelled for the perennial *Arabidopsis halleri* [7]. The influences of temperature and photoperiod on gene regulation were explicitly incorporated in the ODE model, in such a way that the expression of the genes could be simulated for different temperatures. This allowed to forecast changes in the perennial flowering cycle under a climate change scenario. Noteworthy, if the genes and processes that underlie the signal perception are included in the model, similar strategies can be applied to represent the effect of vernalization in our model. In addition to forecasting the dynamics of gene expression under diverse environmental conditions, we believe that such models can aid to explore hypothesis for the molecular mechanisms underlying the perception and integration of external signals. This could be achieved e.g. by using phenotypic information of plants grown under different environmental conditions to fit the data; then by exploring the correlation between changes in external signals with changes in gene expressions and parameters. However this would require more resolution of the experimental data and parameters than is currently available.

Integrative data analysis that bridges the different levels of systems organization

Dynamic modelling is not the only tool for plant systems biology [8]; systems biology also concerns the analysis of the multiple aspects of a system. Instead of using mathematical models, multiple experiments are performed in such a way that bioinformatics tools and biostatistics methods can be applied to analyse the data integratively. Often, these experiments are performed independently (e.g. by different labs), thus adding extra challenges to the analysis such as scaling, normalization and standardization. Yet, as reviewed in Chapter 1, integrative data analysis has been shown a powerful strategy in unveiling properties that increase our understanding of GRNs underlying plant processes. But we are just at the beginning of understanding the different levels of a system's organisation and how they should be integrated to obtain a more global picture of a biological process. An obvious connection is between genetic variation and a GRN. It is evident that polymorphisms among individuals may alter the network and/or the dynamics of the GRNs underlying plant processes [9,10]. In a similar way, environmental factors play an important role in plant development by perturbing the molecular network [1,11,12]. Therefore, it is important to develop a deeper understanding of how these levels of biological organization are interconnected. This will enable us to understand e.g. how

environmental changes drive genetic perturbations and how these perturbations affect the GRNs underlying plant development. Furthermore, we will be able to understand how a perturbation in the GRN is propagated sequentially over the developmental course through changes in the stage-specific states of a GRN. In that direction, in Chapter 3 we were interested in identifying which genetic perturbations may influence the flowering time response, and we were also interested in understanding how they do so. For that, we inquired simultaneously omics data from three levels of systems organization; 1) polymorphisms among Arabidopsis accessions at the population level; 2) flowering time phenotype at an individual level; and 3) the physical and genetic interactions of genes and proteins affected by the polymorphisms at a subcellular level. Based on the identified single nucleotides polymorphisms (SNPs) and on available experimental data, we formulated hypothesis for explaining the molecular mechanisms underlying the effect the polymorphisms on gene regulation of flowering time genes. Interestingly, our method reveals statistical dependencies between the identified polymorphisms. For a few cases, hypotheses for the molecular mechanism are only consistent when these dependencies are taken into account. One future step is to analyse such dependencies. This can be done by e.g. assessing if simultaneous SNPs in co-dependent genes are necessary to alter a phenotype, or if a SNP in a co-dependent gene can override the phenotypic difference caused by a first SNP. Another future step is to use quantitative information about the flowering time phenotype. Currently, we use a binary classification for the phenotype in which the accessions are classified as Early or Late flowering based on comparison against the flowering time of the reference ecotype Columbia. One alternative method that could be used would be based on Random forest [13], which might improve prediction performance compared to our decision tree model.

In Chapter 4, we developed and applied a bioinformatics tool that predicts protein-protein interaction (PPI) sites at a proteome-wide scale. In order to perform the predictions, the tool itself integrates multidisciplinary aspects of interactomics (PPI network), proteomics (the sequences of the proteins and properties calculated from them) and phylogenomics (the conservation of protein sequences across species). We implemented the tool based on the assumption that protein sequence motifs that a) are overrepresented in pairs of interacting proteins, and that b) are highly conserved across orthologs, and that c) are exposed to the surface of the protein structure are good putative protein-protein interaction sites. Importantly, we interrogated Arabidopsis interactome data, together with the predicted protein-protein binding motifs, to formulate testable hypotheses for the molecular mechanisms underlying effects of protein sequence polymorphisms. The hypotheses were formulated for proteins from which the position of the predicted protein-protein binding motif overlaps the position of an annotated mutagenesis site (either natural variation or site-directed mutagenesis mutations). Based on this and on the interactome data, we could determine specifically which PPIs are

being disrupted because of the mutations, and also which interactions are maintained in spite of the mutations. With this information, we then proposed hypotheses to explain the effect of the mutations on the protein interaction specificity and how this is related to the observed change in e.g. expression.

Chapter 4 is another example of integrating and inquiring data from different 'omics' towards understanding the genotype-to-phenotype relationship. The method presented in Chapter 4 uses interactomics (PPI network), proteomics (the sequences of the proteins and properties calculated from them) and phylogenomics (the conservation of protein sequences across species). Specifically, the residue surface accessibility (RSA) is used to assess if an amino-acid is exposed to the surface of the protein structure. We used this approach because for the majority of proteins represented in the Arabidopsis interactome, no 3D structural information is available. Although the overall correlation coefficients between the actual and predicted RSA reaches up to 0.66 [14], we identified a potential for improving our method by using alternative methods that infer the RSA from sequence [15,16] or that calculate the RSA from structure of homologous proteins. Another possibility for improving our method would be by using weighted graphs to represent a PPI network. The weight of an interaction connecting two proteins could represent e.g. the strength of the interaction or the probability of that link being a true positive interaction. To start with, statistical methods for inferring these measures from yeast-two-hybrid assay results could be developed based on the scoring schemes for identifying the interaction. This semi-quantitative information about an interaction would then be incorporated in the algorithm of our method to compute degree of over-representation of a motif in the network. This could be achieved e.g. by statistically combining the measure for the over-representation of the motif with the semi-quantitative information about the interaction. Finally, it would be interesting to analyse the motifs responsible not only for dimerization but also for higher-order complex formation (e.g. tetramers). For that our method could be adapted to mine motifs in those sub graphs from the interactome that represent the connection patterns expected for e.g. MADS domain protein heterotetramers [17]. This could point to multiple motifs in the same protein, representing both the dimerization and heterodimerization sites.

Recently, Pajoro et al. [18] used a combination of DNAseI-seq, ChIP-seq and microarrays to study the link between TF, chromatin and expression during flower development. It was shown that the binding of MADS-box transcription factors (TF) APETALA1 (AP1) and SEPALATA3 (SEP3) precedes opening of chromatin. Based on the results of that study, it was hypothesized that MADS-box TFs are able to act as pioneer factors, i.e. they are able to bind to closed chromatin and then directly or indirectly opens it in order to facilitate the binding of other transcription factors to that particular region. A recent study shows that MADS-box TFs interact with chromatin modifiers and remodelers [6], thereby supporting an indirect effect of MADS-

box TFs on the chromatin. The study of Pajoro et al. shows a potential mechanism by which MADS-box TFs may regulate a set of target gene: by regulating the 'chromatin state' in the regulatory region of their targets.

In general, the chromatin state and the various chromatin modifications are important parameters of gene expression and will be more incorporated into expression and GRN models.

Recently, a combination of ChIP, gene expression and modelling techniques has successfully been applied on the study of the epigenetic mechanism underlying cold perception and memory in *Arabidopsis* [19]. This mechanism, which involves epigenetic silencing of *FLC* gene by the Polycomb repressive complex (PRC1), has been implicated in the control of flowering time [20] In [19], the authors used a previously proposed stochastic modelling framework [21] that takes into account only three possible configurations for the histones; activating (A), unmodified (U) and H3K27me3 (M). The activating (A) state is believed to has its own marks, while the H3K27me3 (M) state represents the repressed state after the activity of PRC. The authors consider the average histone status (A,U or M) of all histones within a specific hotspot region characterized by increasing H3K27me3 (M) during cold exposure. This hotspot region is close to the transcription start site of *FLC* and has remarkable high levels of tri-methylated H3K27 when the plants are exposed to cold. Interestingly, the model is not only able to simulate the experimental patterns of H3K27me3 (M) during and after cold exposure, but it also predicts bi-stable states which correlates with the bi-stable patterns of *FLC* expression during cold exposure. Similar approaches using stochastic modelling framework can be used to study temporal dynamics of epigenetic elements during flower development, especially when data about nucleosome positioning and histone modifications are available.

Connecting the dots

In Chapter 3, we identified single nucleotide polymorphisms (SNPs) that possibly change the binding affinity of TFs to the promoter of the flowering time genes. For that, we focused only on regulatory SNPs, whilst we ignored the non-synonymous SNPs located in the coding region of the TFs. However, we note that the SNPs that overlap the DNA binding domain of the TF also have the potential to change the TF DNA binding affinity. In addition, in particular for the MADS proteins which are known to require complex formation to bind to the DNA [22] and whose combinatorial interactions largely influence the DNA binding specificity [6], SNPs that overlap the protein-protein interaction domain are also important candidates. This because they may change how the TFs interact and hence the gene expression patterns of their targets. In this respect, a comprehensive catalogue of polymorphisms with effect on gene expression of flowering time genes would survey not only for variations in the regulatory region of genes, but also on coding regions that encode for protein-protein interaction sites and protein-DNA

binding sites. For this, transcriptome data of the Arabidopsis accessions would be very helpful to determine the effect of polymorphisms on gene expression. This can be complemented by the interactome-wide predictions of protein-protein interaction binding sites presented in Chapter 4. The idea is to use the predictions presented in Chapter 4 to annotate SNPs that are located in TF functional parts which may alter the gene expression. By doing so, we will be able to identify polymorphisms that disrupt either the regulatory region of flowering time genes or the protein-protein interaction sites of their upstream regulators; in both cases, we would be interested in polymorphisms that result in altered gene expression patterns and altered flowering time response. As recently envisioned by [23], GRN quantitative models can be used to study the effect of polymorphisms on gene expression and the subsequent effect on the flowering time. Basically, the approach would be to simulate the model to quantify the effect of a SNP on the flowering time phenotype; and this could be achieved by varying the values of the parameters that are presumably affected by the SNP.

To conclude, progress in understanding how the genome links to the phenotype, as well as the effect of the many interactions with environment, can be made when performing multidimensional bioinformatics data analysis, while modelling techniques and frameworks offer promising tools to represent and explore the available information. However, we are just at the beginning of a full representation of GRNs underlying the control of developmental processes; and we are just beginning to understand the interplay between TFs, other regulatory proteins and the chromatin. Challenges for the near future will be (a) to unravel the spatial and temporal regulation of the genes in the current networks, (b) to increase the quantitative resolution of the available information and (c) to integrate various levels and types of data into predictive models. An ultimate goal is to include genetic diversity in the analysis to identify relevant polymorphisms explaining the behaviour of the system and the phenotype of the plant.

References

1. Alice Pajoro SB, Evangelia Dougali, Felipe Leal Valentim, Marta Adelina Mendes, Aimone Porri, George Coupland , Yves Van de Peer , Aalt D.J. van Dijk , Lucia Colombo , Brendan Davies and Gerco C. Angenent (2014) The (r)evolution of gene regulatory networks controlling Arabidopsis plant reproduction, a two decades history. Journal of Experimental Botany Accepted for publication.
2. Karlebach G, Shamir R (2008) Modelling and analysis of gene regulatory networks. Nat Rev Mol Cell Biol 9: 770-780.
3. Kim PM, Tidor B (2003) Limitations of quantitative gene regulation models: a case study. Genome Res 13: 2391-2395.
4. Gonze D (2011) Modeling circadian clocks: Roles, advantages, and limitations. Central European Journal of Biology 6: 712-729.
5. Spitz F, Furlong EE (2012) Transcription factors: from enhancer binding to developmental control. Nat Rev Genet 13: 613-626.
6. Smaczniak C, Immink RG, Muino JM, Blanvillain R, Busscher M, et al. (2012) Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. Proc Natl Acad Sci U S A 109: 1560-1565.
7. Satake A, Kawagoe T, Saburi Y, Chiba Y, Sakurai G, et al. (2013) Forecasting flowering phenology under climate warming by modelling the regulatory dynamics of flowering-time genes. Nat Commun 4: 2303.
8. Yuan JS, Galbraith DW, Dai SY, Griffin P, Stewart CN, Jr. (2008) Plant systems biology comes of age. Trends Plant Sci 13: 165-171.
9. de Bruijn S, Angenent GC, Kaufmann K (2012) Plant 'evo-devo' goes genomic: from candidate genes to regulatory networks. Trends Plant Sci 17: 441-447.
10. Rosas U, Mei Y, Xie Q, Banta JA, Zhou RW, et al. (2014) Variation in Arabidopsis flowering time associated with cis-regulatory variation in CONSTANS. Nat Commun 5: 3651.
11. Fornara F, de Montaigu A, Coupland G (2010) SnapShot: Control of flowering in Arabidopsis. Cell 141: 550, 550 e551-552.
12. Reeves PH, Coupland G (2000) Response of plant development to environment: control of flowering by daylength and temperature. Curr Opin Plant Biol 3: 37-42.
13. Kursa MB (2014) Robustness of Random Forest-based gene selection methods. BMC Bioinformatics 15: 8.
14. Adamczak R, Porollo A, Meller J (2005) Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 59: 467-475.
15. Ali SA, Hassan I, Islam A, Ahmad F (2014) A Review of Methods Available to Estimate Solvent-Accessible Surface Areas of Soluble Proteins in the Folded and Unfolded States. Curr Protein Pept Sci.
16. Huynen MA, Snel B, von Mering C, Bork P (2003) Function prediction and protein networks. Curr Opin Cell Biol 15: 191-198.
17. Espinosa-Soto C, Immink RG, Angenent GC, Alvarez-Buylla ER, de Folter S (2014) Tetramer formation in Arabidopsis MADS domain proteins: analysis of a protein-protein interaction network. BMC Syst Biol 8: 9.
18. Pajoro A, Madrigal P, Muino JM, Matus JT, Jin J, et al. (2014) Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. Genome Biol 15: R41.
19. Angel A, Song J, Dean C, Howard M (2011) A Polycomb-based switch underlying quantitative epigenetic memory. Nature 476: 105-108.
20. Muller R, Goodrich J (2011) Sweet memories: epigenetic control in flowering. F1000 Biol Rep 3: 13.
21. Dodd IB, Micheelsen MA, Sneppen K, Thon G (2007) Theoretical analysis of epigenetic cell memory by nucleosome modification. Cell 129: 813-822.

22. Schwarz-Sommer Z, Huijser P, Nacken W, Saedler H, Sommer H (1990) Genetic Control of Flower Development by Homeotic Genes in Antirrhinum majus. Science 250: 931-936.
23. Marjoram P, Zubair A, Nuzhdin SV (2014) Post-GWAS: where next? More samples, more SNPs or more biology? Heredity (Edinb) 112: 79-88.

Summary

Developmental processes are controlled by regulatory networks (GRNs), which are tightly coordinated networks of transcription factors (TFs) that activate and repress gene expression within a spatial and temporal context. In *Arabidopsis thaliana,* the key components and network structures of the GRNs controlling major plant reproduction processes, such as floral transition and floral organ identity specification, have been comprehensively unveiled. This thanks to advances in 'omics' technologies combined with genetic approaches. Yet, because of the multidimensional nature of the data and because of the complexity of the regulatory mechanisms, there is a clear need to analyse these data in such a way that we can understand how TFs control complex traits. The use of mathematical modelling facilitates the representation of the dynamics of a GRN and enables better insight into GRN complexity; while multidimensional data analysis enables the identification of properties that connect different layers from genotype-to-phenotype. Mathematical modelling and multidimensional data analysis are both parts of a systems biology approach, and this thesis presents the application of both types of systems biology approaches to flowering GRNs.

Chapter 1 comprehensively reviews advances in understanding of GRNs underlying plant reproduction processes, as well as mathematical models and multidimensional data analysis approaches to study plant systems biology. As discussed in Chapter 1, an important aspect of understanding these GRNs is how perturbations in one part of the network are transmitted to other parts, and ultimately how this results in changes in phenotype. Given the complexity of recent versions of *Arabidopsis* GRNs - which involves highly-connected, non-linear networks of TFs, microRNAs, movable factors, hormones and chromatin modifying proteins - it is not possible to predict the effect of gene perturbations on e.g. flowering time in an intuitive way by just looking at the network structure. Therefore, mathematical modelling plays an important role in providing a quantitative understanding of GRNs. In addition, aspects of multidimensional data analysis for understanding GRNs underlying plant reproduction are also discussed in the first Chapter. This includes not only the integration of experimental data, e.g. transcriptomics with protein-DNA binding profiling, but also the integration of different types of networks identified by 'omics' approaches, e.g. protein-protein interaction networks and gene regulatory networks.

Chapter 2 describes a mathematical model for representing the dynamics of key genes in the GRN of flowering time control. We modelled with ordinary differential equations (ODEs) the physical interactions and regulatory relationships of a set of core genes controlling *Arabidopsis* flowering time in order to quantitatively analyse the relationship between their expression levels and the flowering time response. We considered a core GRN composed of eight TFs: *SHORT VEGETATIVE PHASE (SVP), FLOWERING LOCUS C (FLC), AGAMOUS-LIKE 24 (AGL24), SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 (SOC1), APETALA1 (AP1), FLOWERING LOCUS T (FT), LEAFY (LFY)* and *FD*. The connections and interactions amongst these components are justified based on experimental data, and the model is parameterised by fitting the equations to quantitative data on gene expression and flowering time. Then the model is validated with transcript data from a range of mutants. We verify that the model is able to describe some quantitative patterns seen in expression data under genetic perturbations, which supported the credibility of the model and its dynamic properties. The proposed model is able to predict the flowering time by assessing changes in the expression of the orchestrator of floral transition *AP1*. Overall, the work presents a framework, which allows addressing how different quantitative inputs are combined into a single quantitative output, i.e. the timing of flowering. The model allowed studying the established genetic regulations, and we discuss in Chapter 5 the steps towards using the proposed framework to zoom in and obtain new insides about the molecular mechanisms underlying the regulations.

Systems biology does not only involve the use of dynamic modelling but also the development of approaches for multidimensional data analysis that are able to integrate multiple levels of systems organization. In Chapter 3, we aimed at comprehensively identifying and characterizing *cis*-regulatory mutations that have an effect on the GRN of flowering time control. By using ChIP-seq data and information about known DNA binding motifs of TFs involved in plant reproduction, we identified single-nucleotide polymorphisms (SNPs) that are highly discriminative in the classification of the flowering time phenotypes. Often, SNPs that overlap the position of experimentally determined binding sites (e.g. by ChIP-seq), are considered putative regulatory SNPs. We showed that regulatory SNPs are difficult to pinpoint among the sea of polymorphisms localized within binding sites determined by ChIP-seq studies. To overcome this, we narrowed the resolution by focusing on the subset of SNPs that are located within ChIP-seq peaks but that are also part of known regulatory motifs. These SNPs were used as input in a classification algorithm that could predict flowering time of Arabidopsis accessions relative to Col-0. Our strategy is able to identify SNPs that have a biological link with changes in flowering time. We then surveyed the literature to formulate hypothesis that explain the regulatory mechanism underlying the difference in phenotype conferred by a SNP. Examples include SNPs that disrupt the flowering time gene *FT*; in which the mutation presumably

disrupts the binding region of *SVP*. In Chapter 5 we discuss the steps towards extending our approach to obtain a more comprehensive survey of variants that have an effect on the flowering time control.

In Chapter 4, we propose a method for genome-wide prediction of protein-protein interaction (PPI) sites form the Arabidopsis interactome. Our method, named SLIDERbio, uses features encoded in the sequence of proteins and their interactions to predict PPI sites. More specifically, our method mines PPI networks to find over-represented sequence motifs in pairs of interacting proteins. In addition, the inter-species conservation of these over-represented motifs, as well as their predicted surface accessibility, are take into account to compute the likelihood of these motifs being located in a PPI site. Our results suggested that motifs overrepresented in pairs of interacting proteins that are conserved across orthologs and that have high predicted surface accessibility, are in general good putative interaction sites. We applied our method to obtain interactome-wide predictions for Arabidopsis proteins. The results were explored to formulate testable hypothesis for the molecular mechanisms underlying effects of spontaneous or induced mutagenesis on e.g. ZEITLUPE, CXIP1 and SHY2 (proteins relevant for flowering time). In addition, we showed that the binding sites are under stronger selective pressure than the overall protein sequence, and that this may be used to link sequence variability to functional divergence.

Finally, Chapter 5 concludes this thesis and describes future perspectives in systems biology applied to the study of GRNs underlying plant reproduction processes. Two key directions are often followed in systems biology: 1) compiling systems-wide snapshots in which the relationships and interactions between the molecules of a system are comprehensively represented; and 2) generating accurate experimental data that can be used as input for the modelling concepts and techniques or multi-dimensional data analysis. Highlighted in Chapter 5 are the limitations in key steps within the systems biology framework applied to GRN studies. In addition, I discussed improvements and extensions that we envision for our model related to the GRN underlying the control of flowering time. Future steps for multi-dimensional data analysis are also discussed. To sum up, I discussed how to connect the different technologies developed in this thesis towards understanding the interplay between the roles of the genes, developmental stages and environmental conditions.

Samenvatting

Ontwikkelingsprocessen worden gecontroleerd door genregulatienetwerken (GRNs): netwerken van transcriptiefactoren (TF's) die genexpressie reguleren. In *Arabidopsis thaliana* zijn de GRNs ontrafeld die reproductie processen, zoals bloeitijd en bloemontwikkeling, reguleren. Dit dankzij de vooruitgang in 'omics' technologie gecombineerd met genetische technieken. Door de multidimensionale aard van de gegevens en vanwege de complexiteit van de regulerende mechanismen maakt de gegevens nog niet meteen duidelijk hóe deze de GRNs zulke complexe eigenschappen reguleren. Het gebruik van wiskundige modellering kan de dynamiek van een GRN beschrijven om zo een beter inzicht te krijgen in de complexiteit van een GRN, terwijl multidimensionale data-analyse de identificatie mogelijk maakt van de eigenschappen die verschillende lagen van genotype-naar-fenotype verbinden. Wiskundige modellering en multidimensionale data-analyse zijn beide delen van een systeembiologische benadering, en dit proefschrift presenteert de toepassing van beide soorten benaderingen met netwerken die bloeiprocessen reguleren.

Hoofdstuk 1 geeft een overzicht van de vooruitgang in het begrijpen van GRNs die reproductie processen in de plant besturen, evenals wiskundige modellen en multidimensionale data-analyse methoden om planten systeembiologisch te bestuderen. Een belangrijk aspect van het begrijpen van GRNs is hoe verstoringen in een deel van het netwerk naar andere delen worden uitgezonden, en uiteindelijk hoe dit resulteert in veranderingen in fenotype. Gezien de complexiteit van recente versies van Arabidopsis GRNs met niet-lineaire netwerken van TFs, microRNA, transporteerbare factoren, hormonen en chromatine modificerende eiwitten - is het niet mogelijk het effect van genetische verstoringen op bijvoorbeeld bloeitijd te voorspellen door alleen te kijken naar de structuur van het netwerk. Daarom speelt wiskundige modellering een belangrijke rol bij het verstrekken van een kwantitatief inzicht in GRNs. Eveneens worden aspecten van multidimensionale gegevensanalyse voor het begrijpen van GRNs ook besproken in het eerste hoofdstuk. Dit omvat niet alleen de integratie van experimentele gegevens, bijvoorbeeld transcriptomics met eiwit-DNA-binding profielen, maar ook de integratie van verschillende soorten netwerken die door 'omics' benaderingen experimenteel worden gekarakteriseerd, zoals eiwit-eiwit interactie netwerken en gen regulerende netwerken.

Hoofdstuk 2 presenteert een wiskundig model voor het beschrijven van de dynamiek van de belangrijkste genen in de GRN van bloeitijd controle. We hebben met differentiaalvergelijkingen (ODE) de interacties beschreven van een set van genen die in Arabidopsis bloeitijd reguleren. Hiermee is de relatie tussen hun expressie niveaus en bloeitijd

kwantitatief te analyseren. We hebben een hoofd-GRN samengesteld uit acht TFs: *SHORT VEGETATIVE PHASE (SVP), FLOWERING LOCUS C (FLC), AGAMOUS-LIKE 24 (AGL24), SUPPRESSOR OF OVEREXPRESSION OF CONSTANS 1 (SOC1), APETALA1 (AP1), FLOWERING LOCUS T (FT), LEAFY (LFY)* en *FD*. De interacties tussen deze componenten zijn gebaseerd op experimentele gegevens, en waarden voor de parameters in het model worden gevonden met behulp van kwantitatieve gegevens over genexpressie en bloeitijd. Vervolgens wordt het model gevalideerd met genexpressiegegevens van verschillende mutanten. We controleren of het model in staat is om een aantal kwantitatieve patronen te zien in expressie data onder genetische verstoringen. Kortom, het werk presenteert een kader dat aangeeft hoe verschillende kwantitatieve gegevens worden gecombineerd tot een kwantitatieve uitkomst, nl. de timing van bloei.

Systeembiologie betreft niet alleen het gebruik van dynamische modellering, maar ook de ontwikkeling van methoden voor multidimensionale gegevensanalyse die meerdere niveaus van systemen integreren. In hoofdstuk 3 richten we ons op het identificeren en karakteriseren van cis-regulerende mutaties, die een effect hebben op de GRN van bloeitijd controle. Door het gebruik van ChIP-seq data en informatie over bekende DNA-bindende motieven van TFs, die betrokken zijn bij reproductie van planten, identificeerden we single-nucleotide polymorphismen (SNPs) die zeer onderscheidend zijn in de indeling van de bloeitijd fenotypes. SNPs die de positie van experimenteel bepaalde bindingsplaatsen (bijvoorbeeld door ChIP-seq) overlappen, kunnen worden beschouwd als mogelijk regulerende SNPs. We toonden aan dat deze SNPs moeilijk zijn aan te wijzen door de enorme hoeveelheid SNPs. Om dit probleem aan te pakken, focussen we op de subset van SNPs die zich binnen ChIP-seq pieken bevinden, maar ook onderdeel zijn van bekende regulatie motieven. Deze SNPs werden gebruikt als input in een classificatie algoritme dat bloeitijd kon voorspellen van Arabidopsis accessies ten opzichte van Col-0. Onze strategie is in staat om SNPs te identificeren die gecorreleerd zijn met veranderingen in het bloei tijdstip. Een voorbeeld hiervan zijn SNPs die het bloeitijd gen FT verstoren, waarbij de mutatie mogelijk het bindingsgebied van SVP verstoort. In hoofdstuk 5 bespreken we de stappen op weg naar de uitbreiding van onze benadering naar een meer volledig overzicht van de varianten die een effect op de bloeitijd controle hebben.

In hoofdstuk 4, stellen we een methode voor die eiwit-eiwit interactie sites voorspelt in het Arabidopsis interactoom. Onze methode, genaamd SLIDERbio, vindt sequentiemotieven die oververtegenwoordigd zijn in paren interagerende eiwitten. Bovendien geeft de methode de motieven die geconserveerd zijn tussen verschillende soorten, en of het waarschijnlijk is dat ze zich bevinden aan het oppervlak van een eiwit. We pasten onze methode toe op het interactoom van Arabidopsis. De resultaten werden onderzocht om toetsbare hypothesen te formuleren voor de moleculaire mechanismen die ten grondslag liggen aan het effect van spontane of

geïnduceerde mutagenese op bv Zeitlupe, CXIP1 en SHY2 (eiwitten die relevant zijn voor de bloei tijd). Bovendien toonden we aan dat de bindingsplaatsen onder sterkere selectie druk staan dan de totale eiwitsequentie; dit kan worden gebruikt om sequentievariabiliteit te verbinden met functionele verschillen.

Tot slot, beschrijft hoofdstuk 5 van dit proefschrift toekomstperspectieven in de systeembiologie toegepast op de studie van GRNs die reproductie processen in de plant reguleren. Twee belangrijke richtingen worden vaak gevolgd in systeembiologie: 1) het opstellen van systemen-brede snapshots waarin de relaties en interacties tussen de moleculen van een systeem worden beschreven; en 2) het genereren van nauwkeurige experimentele gegevens die kunnen worden gebruikt als input voor modelleren en/of multidimensionale gegevensanalyse. In hoofdstuk 5 worden de beperkingen in de belangrijkste stappen in systeembiologie besproken. Daarnaast bespreek ik verbeteringen en uitbreidingen die wij voor ogen hebben voor ons model van de bloeitijd- GRN. Toekomstige stappen voor multi-dimensionale data-analyse worden ook besproken. Kortom, ik bespreek hoe de verschillende technologieën ontwikkeld in dit proefschrift geïntegreerd kunnen worden om het samenspel tussen genen, ontwikkelingsstadia en de omgevingsomstandigheden te begrijpen.

Acknowledgments

I want to start this section by acknowledging the help I have received from my direct supervisor and co-promoter Aalt-Jan van Dijk. Thank you for your infinite patience while teaching systems biology; and for having the doors of your office always opened for sharing your expertise. Also, I enjoyed a lot the dynamics of our collaboration; your scientific pragmatism always balanced my eagerness to risk towards newer and more ambitious results. The scientific merit of this thesis I attribute greatly to your pragmatism; while the possible inaccuracies are my ambition´s fault. I am looking forward to experience how the dynamics of our collaboration will evolve now that we will work independently. Finally, I would like to thank you for you always emanating values that were important for the development of myself not only as a scientist; e.g. hard-work, family-life and helpfulness. I think these are important things to remind young PhD students of.

Next I would like to acknowledge all the support I have received from my promoter Gerco Angenent. First, thank you for the invitation to participate in the Veluweloop race 2012 together with the members of the Plant Developmental Systems (PDS) cluster. It was fun and challenging to run a sport event with a group. Also, thank you for giving me the opportunity to execute my PhD within the SYSFLO network. I enjoyed all the travels, meetings, dinners and also to prepare for the several presentations. In the same direction, thank you for the opportunity to explore possibilities within and to participate in the PDS cluster. It was motivating to get involved with such smart, motivated and open group of people. My expertise developed during my PhD should be on Systems Biology and Bioinformatics but I (still) read with great interest and enthusiasm about flowering time regulation because of my participation within these groups; I am grateful for that. Additionally, thank you for the several meetings along the course of my PhD. Not only did I enjoy the scientific discussions but I also tried to learn diverse aspects from all those encounters. Finally, I would like to thank you for, in the beginning of PhD, encouraging my optimism towards achieving goals; and in the end of PhD, for helping me towards finishing the thesis within a quick time frame.

I would like to thank Richard Immink for all the scientific input about MADS-box proteins, gene regulation and regulatory networks. Thank you for your consistent willingness to share scientific information and knowledge; this was very relevant (if not essential) for my understanding of the gene regulatory network of flowering time control. Thank you also for

giving me the opportunity to participate in the group effort for analysing and interpreting the ChIP-seq of SOC1; I did enjoy my role in that work. Finally, thank you for the visit when I got ill; I am grateful for that.

I also would like to thank Roeland van Ham for giving me the opportunity to start the PhD in -back in 2010- his group. His leadership style allowed me to focus completely on the scientific aspects of my thesis; and in retrospect, I can recognize exponential growth on my scientific capabilities as result of our collaboration during the first year and a half of my PhD. Besides, he posed powerful questions that I still try to answer whenever I have to define my professional role within a scientific project or network; thank you for that.

Very important for the accomplishment of this PhD and for my development as a scientist was my participation in the SYSFLO network. I am grateful to all the PhD fellows: Sandra Biewers, Alice Pajoro, Marta Mendes, Aimone Porri, Katharina Schiessl, Miguel Godinho, Pedro Madrigal Bayonas, Evangelia Dougali; to all the professors: Brendan Davies, Gerco Angenent, Lucia Colombo, George Coupland, Robert Sablowski, Pawel Krajewski and Yves Van de Peer; to the principal investigators Aalt-Jan van Dijk and Birgit Lewicki-Potapov; and the project manager Juliet Jopson. Thank you for all the interactions, feedbacks, meetings, dinners, discussions and the good time. Special thanks to Aimone Porri and George Coupland for having received me in the Max Planck Institute Köln. The time I spent visiting the MPI in Köln added substantially to my formation, and I did enjoy my time in MPI Köln. I also acknowledge the essential role of the Marie Curie Actions by supporting the SYSFLO network and the training of the PhD fellows.

In addition, also very important for the accomplishment of this PhD was the collaboration with Peter Boyen and Frank Neven, from Hasselt University. I hope we can plan and execute more projects as smoothly as we did for the SLIDERBio. And also, the collaboration for the quantitative modelling project; Simon van Mourik, David Posé, Markus Schmid, Marcos Busscher and Jaap Molenaar. Thank you all for sharing data, expertise, motivation and ideas. I also acknowledge the help I received from Antonio Chalfun Junior, Andre Saude and Luciano Vilela Paiva; their support during the MSc was essential for me to develop interest for plant molecular biology, and in the end of my PhD, they allowed me to use their working environment to conclude my thesis; thank you for that.

I would like to thank also some good friends I have made in Wageningen; for all the interesting dinners, good talks, crazy ideas and small adventures. In addition to their friendship, they often assumed family-member roles in my life in Wageningen; thank you Padraic Flood, Natalia Carreño, Charles Neris Moreira, Julio Maia, Veronika Wehner and Ilias Rotsias. A special thanks to Julio, who has been in touch since the time back in Lavras/Brazil; and Veronika, for

the unconditional support and friendship. Then, I would like to thank others PhD students who, each one by his own way, contributed to my PhD formation and personal experience in Wageningen; thank you Erik Wijnker, Alice Pajoro, Cezary Smaczniak, Anneke Horstman, Leonie Verhage, Etalo Desalegn, Martin Kompmaker, Renake Teixeira, Jennifer de Jonge, Kalaitzoglou Pavlos and others. A special thanks goes to my former office mates Saulo Aflitos and Joachim Bargsten – Joachim, I wish all the bests for your newly formed family; Saulo, whenever you decide to get married, please invite me. Finally, I would like to acknowledge the pleasant time I spent collaborating with the two (back then) master students; Laurie Thibaudat and Mathilde Botineu.

I am especially grateful to the Applied Bioinformatics group of the Plant Research International; for all the scientific and personal support. Thank you for all the interactions, coffee breaks, literature and work discussions. Thank you for all the direct, Dutch style, feedback. I truly enjoyed greatly my time in Wageningen mostly because of my participation in the team. It was fun and challenging to discuss science with you during my PhD: Jose Muiño, Joachim Bargsten, Jan Peter Nap, Gabino Sanchez Perez, Jan van Haarst, Aalt-Jan van Dijk, Henri van de Geest, Bas Lintel, Hekkert, Edouard Severing, Saulo Alves Aflitos, Pieter Lukasse, Sander Peters, Elio Schijlen, Sandra Smit, Ke Lin and Roeland van Ham. I hope I will have the opportunity to collaborate with you all; and I miss having the opportunity for a proper farewell.

I have no words to express my gratitude towards my father, Valter Alvares Valentim, and my mother Vania Mara Leal Valentim. I can very well remember a myriad of reasons why I should dedicate this thesis to you; but it would be too long and too complicated to do so. To make it short: thank you, I love you and this thesis is dedicated to you. Also, I would like to thank my sister Julia Leal Valentim. Thank you for being my sister and thank you for bringing my nephew Ian to this world. The thoughts, memories and interactions I have about you two give depth to my life. And to my brother Ivan Leal Valentim. Thank you for being this serene, constant person in our family. I am proud of your life trajectory. Finally, my uncle Elson Magalhães Toledo. Thank you for all the sincere, personal, supportive and sometimes funny e-mails. They made me reflect about the rest of our family - I am grateful for that. And, since you showed interest in the modelling aspects of my thesis, it would be fun to collaborate with you in the LNCC (National Laboratory of Scientific Computing) in the future.

I am also grateful for to Andries Koops, Gabino Sanchez Perez and Ernst van den Ende, Sydney Leijenhorst, Henk Gerritsen, Jacco Zwagerman, Martien Opdam, Thirza V. and Philip van Der Stelt. From Brazil I am grateful to the social engagement in the philanthropic institute José de Arimatea (this embodied specially by Hermane Godoy).

I am sorry if I let any name out of this section. Overall, I would like to thank all the PhD students and researches that I have met in Europe during courses, conferences and workshops. Thank you all for, along the way, exchanging your motivation and experience!

Curriculum Vitae

I received my first academic degree in Computer Science. Because of my interest in bioinformatics I migrated to a Master's degree in Plant Biotechnology. During the PhD I studied Systems Biology and Bioinformatics applied to plant science. I applied systems biology to understand the molecular mechanisms underlying the regulation of flowering-time control. My interest is in developing theories and models to help understanding the molecular mechanisms underlying regulation of complex biological traits. My motivation came from exchanging thoughts with biologists about these biological traits. During my PhD and MSc degrees I have acquired experience with computational biology, systems biology and computational genomics.

Education Statement of the Graduate School
Experimental Plant Sciences

Issued to: **Felipe Leal Valentim**

Group: **Molecular Biology & PRI-Bioscience, Wageningen University and Research Centre**

1) Start-up phase	date	cp
► **First presentation of your project**		
SYSFLO: Initial Network Meeting, Castellammare - Quantitative Modelling of gene regulatory network of flowering time control	May 2010	1.5
► **Writing or rewriting a project proposal**		
SYSFLO Project - Systems Biology Applied to Flowering Time	Sep 2010	6.0
► **Writing a review or book chapter**		
The (r)evolution of gene regulatory networks controlling Arabidopsis plant reproduction, a two decades history	Jan 2014	6.0
► **MSc courses**		
► **Laboratory use of isotopes**		
Subtotal Start-up Phase		13.5

2) Scientific Exposure	date	cp
► **EPS PhD student days**		
EPS PhD student day, Wageningen University, Wageningen, NL	May 20, 2011	0.3
3rd European Retreat of EPS PhD Students in Plant Sciences, Orsay, France	Jul 05-08, 2011	1.2
EPS PhD student day, University of Amsterdam, Amsterdam, NL	Nov 30, 2012	0.3
► **EPS theme symposia**		
EPS theme symposium "Developmental Biology of Plants" - Leiden	Jan 20, 2011	0.3
EPS theme symposium "Developmental Biology of Plants" - Leiden	Jan 17, 2013	0.3
► **NWO Lunteren days and other National Platforms**		
Systems Biology Day – Wageningen University	Jun 16, 2010	0.3
NBIC Conference 2011. The 6th edition of the Netherlands Bioinformatics Centre Conference, Lunteren, The Netherlands	Apr 19-20, 2011	0.6
Plant Genome Evolution - Elsevier Current Opinion Conference, Amsterdam, The Netherlands	Sep 06, 2011	0.3
► **Seminars (series), workshops and symposia**		
NCSB 2010 Symposium - Soesterberg, The Netherlands	Oct 21-22, 2010	0.6
NCSB 2011 Symposium - Soesterberg, The Netherlands	0ct 31-Nov 01, 2011	0.6
► **Seminar plus**		
► **International symposia and congresses**		
Plant Morphodynamics Workshop – John Innes Centre, Norwich, UK	Sep 27-29, 2010	0.6
Conference Comparative and Regulatory Genomics in Plants, Gent, Belgium	Apr 11-12, 2011	0.6
Workshop on Molecular Mechanisms Controlling Flower Development, Maratea, Italy	Jun 13-14, 2011	0.6
EMBL Symposia 2011 - Structure and Dynamics of Protein Networks, Heidelberg, Germany	Oct 13-16, 2011	1.2
► **Presentations**		
F. Leal-Valentim (2010) SYSFLO: Initial Network Meeting, Castellammare. Oral Presentation: "Modelling flowering time gene regulatory network"	May 27, 2010	1.0
F. Leal-Valentim (2010) Plant Morphodynamics Workshop – John Innes Centre, Norwich. Oral Presentation: "Modelling flowering time gene regulatory network"	Sep 28, 2010	1.0
F. Leal-Valentim (2010) 3rd International PhD School on Plant Development – University of Regensburg, Retzbach-Würzburg, Germany, Oral Presentation: "Modelling flowering time gene regulatory network"	Oct 08, 2010	1.0
F. Leal-Valentim (2010) NCSB 2010 Symposium - Soesterberg, The Netherlands. Poster: "Quantitative modelling of the regulatory network of flowering time genes"	Oct 21-22, 2010	1.0
F. Leal-Valentim (2011) SYSFLO: 2nd Network Meeting, Berlin. Oral Presentation: "Predicting protein-protein binding sites from Arabidopsis Interactome"	Feb 09, 2011	1.0
F. Leal-Valentim, A. van Dijk (2011) NBIC Conference 2011. The 6th edition of the Netherlands Bioinformatics Centre Conference. Lunteren, The Netherlands. Oral Presentation:"Modelling flowering time gene regulatory network"	Apr 20, 2011	1.0
F. Leal-Valentim (2011) SYSFLO: Mid Term Review Meeting, Maratea, Oral Presentation: "Modelling flowering time gene regulatory network"	Jun 13, 2011	1.0
F. Leal-Valentim (2011) Workshop on Molecular Mechanisms Controlling Flower Development, Maratea, Poster: "Quantitative modelling of the regulatory network of flowering time genes"	Jun 13-14, 2011	re-use
F. Leal-Valentim (2011) EMBL Symposia 2011 - Structure and Dynamics of Protein Networks- Heidelberg, Germany. Poster: "Genome-wide scale prediction of protein interaction motifs using evolutionary information encoded in sequences and interactome networks"	Oct 13-16, 2011	1.0
F. Leal-Valentim (2011) NCSB 2011 Symposium - Soesterberg, The Netherlands. Poster: "Genome-wide scale prediction of protein interaction motifs using evolutionary information encoded in sequences and	Oct 31-Nov 01, 2011	re-use
F. Leal-Valentim (2012) YISB workshop 2012, Bridging the communication gap in systems biology, Arnhem, The Netherlands	Mar 01-03, 2012	1.0
F. Leal-Valentim (2012) 5th International Ph.D. School in Plant Development. Certosa di Pontignano, Siena, Italy. Oral Presentation: "Quantitative modelling of the gene regulatory network of flowering time control"	Sep 25, 2012	1.0
► **IAB interview**		
► **Excursions**		
Subtotal Scientific Exposure		17.8

CONTINUED ON NEXT PAGE

3) In-Depth Studies	date	cp
► **EPS courses or other PhD courses**		
Optimization Techiniques in Bioinformatics and Systems Biology	May 17-21, 2010	1.5
Postgraduate course 'The Art of Modelling'	Aug 23-Sep 03, 2010	3.0
3rd International Ph.D. School on Plant Development	Oct 06-08, 2010	0.9
5th International Ph.D. School in Plant Development	Sep 25-28, 2012	1.2
► **Journal club**		
Literature discussion at Applied Bioinformatics Group.	Mar 2010-Aug 2013	2.5
► **Individual research training**		
MPI for Plant Breeding Research (George Coupland's Group), Köln, Germany	Apr 2012	0.3
Subtotal In-Depth Studies		9.4

4) Personal development	date	cp
► **Skill training courses**		
ESR Training Course on Getting published and Completing your PhD, University of Leeds, UK	Sep 07-09, 2011	0.6
ESR Training Course on Gene Regulation Analysis Tools and IP / Commercialization, at BIOBASE, Wolfenbüttel, Germany	Oct 25-26, 2010	0.6
SYSFLO ESR Training Course on Grant Writing, Grenoble, France.	Feb 03, 2012	0.3
► **Organisation of PhD students day, course or conference**		
Course on Systems Biology - Lecturer of Master class "Quantitative modelling of gene regulatory network of flowering time control"	Oct 10, 2011	0.8
Course on BIF-30806 Advanced Bioinformatics - Lecturer of Master class "Protein Interaction site prediction"	Feb 2012	0.8
► **Membership of Board, Committee or PhD council**		
Subtotal Personal Development		3.1

TOTAL NUMBER OF CREDIT POINTS*	43.8
Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS	
** A credit represents a normative study load of 28 hours of study.*	